'Influence of Clustering on Statistical Analysis in Orthodontic Literature'

Thesis submitted in accordance with the requirements of the University of Liverpool for the Degree of Doctor of Dental Science(Orthodontics)

by

Balraj Kaur Badeshae

July 2020

Acknowledgments

I would like to express my greatest gratitude to both my research supervisors, Dr Burnside (GB) and Dr Flannigan (NF) for their continuous support at every stage in my research project. Without their enthusiastic encouragement and useful feedback, completion of this project would not have been possible.

I am sincerely grateful to Dr Burnside for his time and advise on statistical analysis of the included articles in this research. His expertise and statistical knowledge have been invaluable.

I would like to thank my mum and siblings for their financial support and encouragement throughout my whole study period. Thank you for having faith in me and standing by my side whenever I needed them.

Finally, and most importantly, I wish to thank my husband for his endless love, support and patience by my side in this research journey.

Table of Contents

Abstract

Aim:

To assess the degree by which clustered study designs are correctly addressed in the statistical analysis in the three major orthodontic journals.

Study design:

This was a retrospective, observational study looking at orthodontic articles published in three major orthodontic journals in 2016 and 2017.

Data source:

The contents of the issues of American Journals of Orthodontics and Dentofacial Orthopedics (AJODO), Angle Orthodontist (AO) and European Journal of Orthodontics (EJO) published in 2016 and 2017 were hand searched.

Review method:

Eligible articles were shortlisted by fist author (BB). Articles presenting with clustering effects were identified and whether the clustering effects were accounted for in the statistical analysis were assessed. They were categorised according to either accounted for clustering, non-accounted for clustering or articles with separate analyses of the outcomes. Additionally, information was collected on journal of publication, continent of authorship, type of study, single or multicentre study, collaboration with statistician, number of authors in the article, sample size calculation, cluster type, statistical significance and statistical method used.

Results:

From the 913 articles identified, after exclusion, 162 articles were considered to have clustering effects in the study design. Of the 162 articles, 84(51.9%) articles correctly accounted for the clustering effects, 36 (22.2%) ignored the clustering effects and 42(25.9%) articles had separate analyses done. Mixed model was the most frequently (100%) used statistical method in articles indicated accounting for the clustering effects. The kappa score for intra-examiner reliability was 0.913 indicant an excellent reliability during data extraction. Involvement of statistician was noted as a significant predictor of accounting for clustering effects. Studies involving a statistician have higher odds of accounting of clustering effects. (When including articles with separate analyses: adjusted OR= 2.91, CI= 1.19-7.11, P= 0.019, When excluding articles with separate analyses: adjusted OR= 8.20, CI= 1.65-40.83, P= 0.010)

Conclusion:

Clustering effects are commonly encountered in orthodontic journals especially in relation to multiple site observations within patients or multiple observations collected at multiple time points. In contrast to the study conducted by Koletsi et al, it can be noted that there has been an increase in the percentage of articles accounting for the clustering effects in the statistical analysis. From only 25% of the included articles searched from 2010 and backwards to 51.9% of the included articles published in 2016 and 2017. It is advisable to involve a statistician in a cluster study to ensure the methodological and statistical issues are addressed appropriately.

List of abbreviations

AJODO	American Journal of Orthodontics and Dentofacial Orthopedics
EJO	European Journal of Orthodontics
AO	The Angle Orthodontist
RCT	Randomised clinical trial
ρ	Intracluster correlation coefficient
k	Coefficient of variation
DE	Design effect
ANNOVA	Analysis of variance
X^2	Chi- square test
SJR	SCImago Journal and Country Rank
CONSORT	Consolidated Standards of Reporting Trials
CIs	Confidence intervals
OR	Odd ratio
GEE	Generalised estimating equations

List of Tables:

List of Figures:

Chapter 1: Introduction

Clusters are defined as 'aggregates of individuals/subjects or a collection of multiple measurements belonging to the same subject'. [1, 2] Clinical trials where groups of individuals or clusters are randomised to receive the same treatment are known as cluster randomised trials.[3] [2, 4-6] [7]Clustered data is seen in various types of study such as in cluster randomised trials where research participants or units are allocated to an intervention as a group, in longitudinal studies where repeated measurements are taken from the same individual at multiple time points and in observational studies where an outcome may be analysed at multiple sites within an individual subject.[8]

In a conventional non-cluster randomised trial with two treatment arms, each subject is individually assigned at random to either one of the treatment arms.[9] In these trials, the intervention is applied directly to the individual subject and observations of each individual subject determine the outcome of the intervention.[10] The methods for the design and analysis of such trials are rather well known. However, in a cluster trial, subjects/units are allocated to an either one of the treatment arms in a group rather than independently.[4] These groups of subjects/individuals are referred to as *clusters* and trials which allocate groups randomly to either one of the treatment arms are known as cluster randomised trials.[4, 9]

The main implication of a cluster design trial is that response from individuals within a cluster are likely to be more similar than those from different clusters.[1, 4, 11, 12]This is because individuals within a cluster may share similar characteristics or are exposed to the same external factors associated with the cluster. The lack of independence of the subjects introduces complexity to the study design and requires modifications in the statistical analysis.[3]The degree of similarity or clustering is commonly quantified by the Intracluster Correlation Coefficient (ICC). This similarity within clusters reduces the amount of information obtained compared to observations obtained without clustering. Hence, the sample size required in a clustered design is generally larger compared to an individually randomised trial. The 'design effect'(DE) can be used to estimate the extent to which the sample size should be inflated to accommodate for the similarity of this clustered data.[8] The precise effect of cluster randomisation on sample size requirements depends on both the size of the cluster and the degree of within-cluster dependence.[13]

Additionally, clustered designs require appropriate statistical analyses to account for the fact that observations within clusters may be more similar.[6, 13, 14] Data from studies with clustering effects can

be analysed either at the cluster level, with the cluster as the unit of analysis, or at the individual level accounting for clustering.[8] Failure to account for clustering can lead to inaccurate results and conclusions. [1, 8, 9]

In essence, clinical trials with clustering effects will require additional consideration in the methodology of the trial, sample size consideration and statistical methods when analysing the data. In 2009, Gibson and Harrison investigated the types of study published in four main orthodontic journals- the American Journal of Orthodontics and Dentofacial Orthopedics (AJODO), The Angle Orthodontist (AO), the European Journal of Orthodontics (EJO) and the Journal of Orthodontics (JO) between 1999 and 2008. They found that 75% of the clinical based studies were mainly examining diagnosis, development and treatment of human subjects.[15] Potentially a number of these clinical studies could have used a clustered study methodology. However, is the effect of clustering taken into account in the study design, statistical analysis and interpretation of published orthodontic studies? In the study by Koleksi et al (2011), that included the most recent 24 issues in 3 major orthodontic journals prior to December 2010, a conclusion was made that only 25.20 percent of the included orthodontics studies where clustering was evident, had taken into account the clustering effects during statistical analysis.[1] In this study, of the 250 articles which were considered to have clustering effects, only 63 articles accounted for clustering in the data analysis.[1]

Evidence based dentistry (EBD) is important in the practice of dentistry. The clinician needs to search for the best available evidence and be able to critically appraise the evidence to apply it to the clinical situation appropriately.[16] Therefore, clinicians must always be careful in appraising the published journals as there is a possibility that an inappropriate use of statistical analysis affects the results of the study. This study aims to search the published orthodontic literature for studies presenting with a clustering design and to explore if the clustering effect is taken into account in the data analysis. Furthermore, a comparison with the previous study done by Koleksi et al (2011) is drawn to further assess if there is a change in the proportion of orthodontic studies which account for clustering in statistical analysis. Associations between study characteristics and appropriate statistical analysis are investigated.

Chapter 2: Literature Review

2.1 Basic Concept

2.1.1 Types of clusters:

There are various types of cluster which can be formed in a cluster trial. Systematically, the types of clusters can be classified to the following categories:

I. Geographical clusters

These are clusters demarcated based on geographical areas.[17] This is appropriate for trials of interventions directed at entire populations or subgroup of populations. This includes well defined communities, such as villages or towns and administrative units such as districts or regions.[17]

II. Institutional clusters

This is grouping of participants based on the specific institutions or organisations they belong to such as schools, universities, health units, communities or work places.[17, 18] Health units that are commonly randomised in cluster randomised trials include hospitals, clinics, general practices and individual practitioners. In these cluster trials, the patients attending the health units generally form a cluster. Cluster randomised trials are often applied to evaluate nontherapeutic interventions, including lifestyle modifications, educational programmes and innovations of the provision of health care.[19]

III. Smaller clusters

Smaller groups such as households or families can be considered as a unit of randomisation. They provide logistical convenience and prevents contamination that can occur if different members of families were to be randomised to different treatment arms.[17]

IV. Individual as cluster

Individuals themselves can be considered as clusters.[17] This is the most common type of cluster in dental trials. For example, a clinical study examining teeth surfaces affected by dental caries can generate data with multiple observations in each participants. Clustering effects should be considered when the outcome is the observation on an individual tooth or surfaces.[7] Hence, the oral cavity can be considered a cluster, consisting of several individual teeth.

Additionally, when repeated measurements are carried out on the same individual over time, the observations for each individual can be regarded as a 'cluster'.

2.1.2 Rationale for conducting cluster trials

When designing a cluster RCT, there must be a justifiable rationale for adopting the design.
Following are the reasons for designing a cluster trial:

I. Type of intervention[17]

A cluster trial is suitable where the intervention itself is designed to be delivered to a group rather than to individual subjects.[3, 4] An intervention is best delivered in a group if subjects in a trial cannot be allocated independently or when subjects may interact with one another during the trial. These groups are referred to as clusters. These groups/units may be, children in a class, communities, members of a family, medical practices or teeth in a mouth of an individual patient.[3]

For example, in the assessment of health care strategies, the medical practices or even communities are assigned to the intervention or control group. Here, each medical practice or community forms an institutional cluster and the individuals attending the practice are part of the cluster. In dentistry, clustering is commonly encountered, as the dentition is comprised of multiple teeth, quadrants and jaws.[2] Here, the patient can be considered as a cluster and the teeth present within the oral cavity of each patient can be considered as subjects within the cluster.

II. Contamination[17, 20 21]

In an individually randomised trial, contamination can occur when individuals in one treatment arm receives part or all of the intervention allocated to the other treatment arm. Possibilities of contamination is likely to decrease as cluster size increases or selecting clusters that are well separated.[17]For example, if individual students of the same classrooms are allocated to different treatment groups, it is likely information may be shared with the others in the control group. This is particularly obvious when educational strategies are delivered to intact classrooms of students. As a result, the outcome differences between the treatment arms will be weak.[17] This will bias the trial towards smaller effect estimates.[17] Thus, randomising the students by respective classrooms may be more appropriate, where each classroom acts as a cluster.[3, 17]

III. Logistical convenience and acceptability[17]

A cluster study design is suitable where there are logistic or administrative problems in delivering the intervention to an individual.[3, 17, 20]This is highlighted when randomising general practices in a trial aiming to evaluate the effectiveness of behavioural intervention to lower smoking rates. Two intervention groups are formed; one to offer health promotion using behavioural approach by specially trained practice nurses and the other to use usual general practice care. The outcome measured would be smoking rates in patients from each practice a year later. In this study, randomisation of practices is convenient and cost effective because training of staff at only one practice is required and it also helps to prevent contamination that may occur if individual patients are randomised.

IV. Multiple Measurements from each subject[2, 3]

Clustering is also present when multiple measurements from each subject/individual are being made over time and situations where multiple body parts are being assessed in an individual.[3, 17]As the measurements are done on the same subject/individual, the observations will be correlated and is therefore more appropriate to be treated as a cluster.[22] As an example, in a study aiming to determine the stability of lower labial segment following orthodontic treatment, subjects are followed up for few years and Little's irregularity index in measured at each pre-determined point. Measurements collected from each participant at each time point are regarded as a cluster.

2.1.3 Clustering in orthodontics

Clustering is a common feature of clinical orthodontic research. The following are some common types of study which present with clustering effects:

i. Multiple site observations within the same patient[2]

Clustering is present in studies when observations of multiple sites are collected within a patient.[3] Clustered data are often found in orthodontics when outcomes at level of teeth, sextants, quadrants or jaws are used.[1, 3] This includes assessment of caries, observations of severity of enamel decalcification, assessment of bond failures when on fixed appliance, measurements on alveolar bone thickness, plaque/gingival indexes and even assessment of bilaterally impacted canines. Teeth nested within the same individual are likely to respond more similarly due to the correlated nature and exposure to the similar oral environment. The data can be influenced by several patient-related factors such as patient age, masticatory forces, smoking habit, oral hygiene, compliance, genetic predisposition or systemic disease.

Hence, the data collected from each cluster which is the individual should be well distinguished from data that are collected from another individual.

In these studies, there can be a variety of unit of analysis such as the tooth, the quadrant, surfaces of the tooth, or the individuals' mouth.[3] These units of analysis can be measured as a single summary of measurements of all teeth within the individuals' mouth or as multiple data from each tooth nested in the individuals' mouth.[23] Regardless of the way it has been measured, the data are not independent of each other as they are derived from the same individual and this needs to be accounted for in the analysis.[3]

For example, Bazarghani et al 2016, conducted a cross-mouth RCT, aimed to evaluate the effects of primer on the bracket failure rate in orthodontic patients. In this trial, the outcome measure was based on the number of bracket failure over the study duration. In each patient two diagonal quadrants were randomly assigned to the primer group and contralateral diagonal quadrants to the non-primer group. Therefore, each participant served as their own control. As there are several teeth nested within a patient's mouth, the patient is then considered as a cluster. The presence of clustering can be identified by comparing the total number of observation sites and size of the sample in the study. For example in this study, a total of 908 brackets were assessed in 49 patients. Therefore, the total number of teeth per treatment arm would be number of teeth nested within the diagonal quadrants in each individual multiplied by the total number of patients recruited in the study. 454 brackets were in each treatment arm respectively.[24] However, it would be an error to disregard the fact that each group of teeth from each treatment arm constitutes a cluster as the measures belong to the same patient. Variation between patients represents the between cluster variability, which can influence the rate of bracket failure. Patients with good oral hygiene and dietary habits may potentially have less bracket failure compared to patients with poor oral hygiene and dietary habits. Thus, correlation structure within patients should be considered when evaluating bracket failure rate. Ignoring the clustered nature of the data and treating the individual teeth as independent increases the chance of getting a significant difference which may be misleading. The author of this study conducted an individual level analysis using logistic regression for repeated measurements using generalised estimating equations, with an exchangeable correlation structure within patients to evaluate the bracket failure rate difference between the treatment arms. [24]This was found to be an appropriate statistical method for this study.

ii. Repeated measures collected at consecutive time points

Clustering is also considered in orthodontics when multiple measurements are taken form the same patient over time.[3] This is encountered in trials where the researcher would like to access the changes in the outcome over time. For example, in a study of the effect of fluoride on caries, caries reduction data is collected at the predetermined time points and patients are followed up over a certain period. Measurements from the same individual/subject are expected to be more similar compared to measurements between individuals. This similarity of the within participants creates clustering effect which should be taken into account in the study design and data analysis. It the outcome from all time points are to be included in a single analysis, then the individual can be considered to be a cluster. However, it is also possible to analyse the outcomes from individual time points separately. In this second approach, researcher is unable to access the potential changes over time within an individual and requires additional numbers of statistical tests to be done when analysing the outcomes at each time points.

As an example, Qamruddin et al 2016, conducted a single blinded split mouth controlled trial to determine the effect of a single dose of low-level laser therapy on spontaneous and chewing pain after the placement of elastomeric separators. In this trial, elastomeric separators were placed on either side of the lower molars in all quadrants. The experimental side was treated by low level laser therapy and the opposite side received placebo laser therapy. A numeric rating scale was used to assess the intensity for pain on the next seven days. Pain was the observation in this study and it was reported by each participants at 7 pre-determined time points.[25] Pain threshold and the ratings between patients are different but the responses belonging to the same patient are likely to be correlated. Therefore, the seven different pain scores collected from each individual over the seven days can be considered as a cluster since they represent a collection of measurement belonging to the same individual.

iii. Institutional Cluster

Generally, groups of individuals such as patients in a practice or hospital can be called clusters. Cluster randomisation is used in multicentre trials as patients attending the same hospital may have interacted with one another. These trials are commonly treated as cluster trials due to practical reasons including ethical concerns, financial concerns and the need to avoid treatment group contamination. For example, Mandall et al conducted a multicentre randomised control trial with the aim to investigate the effectiveness of early Class III

protraction facemask treatment in children under 10 years of age. Eight UK hospital orthodontic units were included in this trial. Patients who fulfilled the inclusion criteria were included in this study and randomised to either the protraction facemask or control group. Data was collected to assess the dentofacial changes, occlusal changes, self-esteem, psychosocial impact of malocclusion and Temporomandibular joint (TMJ) symptoms.[26] Data was collected at the start of treatment and 15 months later.[26] In this study, the hospital at the institutional cluster and patients attending at this hospital are the subjects nested within the same cluster.

iv. Multiple Assessors

When there are multiple assessors rating the same image, the end result will include multiple scores of the same photographs. The outcomes will be correlated with each other as they belong to the same photographs. Therefore, these correlations should be taken into account during statistical analysis. This is commonly seen in silhouette studie, where an average value of the score is recorded when taking the clustering effect into account.

Tisouli et al, 2017, investigated on the perceived facial changes in Class II Division 1 patients with convex profiles after functional orthopaedic treatment combined with fixed orthodontic appliances. In this study, profile photographs of pre-treatment and post-treatment of patients treated with activators, twin block appliance and a group of control patients were assessed. There were 12 patients in the respective groups. The photographs were presented in pairs and rated by 3 different group of people, consisting of orthodontists, lay people, parents and patients. There were 10 rates each in respective groups.[27] Therefore, each of the photographs were rated 10 times by each group and 40 times in the whole study. Since each patients was rated by 10 members of each group of raters, the median score was used to obtain a more respective approximation of each group's assessment for each patient.

However, are these clustering effects taken into account in the statistical analysis? Results from the study by Koletsi et al, was rather disappointing where three quarter of the studies with clustering effects evident, did not take these effects into account during data analysis.[1] Another example, Mandall et al, 2003, presented a cochrane review aimed to evaluate the effectiveness of different adhesives for fixed orthodontics brackets. In this review, 25 studies were identified. However, ten of them were excluded due to inappropriate statistical analysis. This was because data analysis was based on the number of bond failure by tooth rather than on a patient basis or included multiple failures per tooth.[28]

2.2 Quantifying the effect of clustering

2.2.1 Variability between clusters

In a conventional clinical trial, individual subjects are randomly allocated to either one of the treatment arms and the characteristics of the individuals are independent of one another. Therefore, standard statistical tests are suitable for this design.[4] However, as discussed in a cluster trials, groups or clusters of individuals are allocated to treatment arms. Therefore, the similar assumption as a conventional clinical trial is usually invalid because the responses from individuals of the same cluster are likely to be more similar.[4, 17] This lack of independence introduces complexity to the design and analysis of cluster RCT.

In cluster randomised trials, there is a positive within-cluster correlation due to variability in the underlying, means of outcome between clusters. If a cluster has different mean response levels, it follows that subjects in the same cluster will tend to have responses that are more similar to each other compared to responses of subjects in different clusters.[17] These between cluster variability and within-cluster correlations should be considered in the designing and analysis of cluster RCT.[29]

The possible reasons for between-cluster variation include the following factors:

a) Subject selections[20, 30]

   This is where the individuals have the choice to choose the cluster which they would like to belong to.[30] This may result in confounding individual level characteristics with membership in particular clusters.[20]

b) Influence of covariates at the cluster level[20, 30]

   This occurs when all individuals in a cluster are affected in a similar manner as a result of exposure to a common environment.[30] For example, in a trial involving medical practices, the characteristics of the providers of the intervention may be related to the outcome measured on the subjects in the cluster. Thus, the between-cluster variation reflects the variation in responses of individual practitioners.[20] In dentistry, individual teeth within one patients mouth would respond similarly compared to the other patients.

c) Effect of personal interactions among cluster members who receive the same interventions[20,30]

Individual within clusters frequently interact and therefore tend to respond similarly [30] For example, community members may discuss their opinions of health education messages, leading to similarities in risk behaviour between members of same community. Likewise, in a trial of interventions against infectious disease, individuals may have effects on transmission to other individuals in the same community.

The primary implication of adopting a cluster randomised design is that outcomes on individuals within the same cluster tend to be correlated. This further results in reduced power in tests of significance, larger standard of errors and wider confidence intervals. [30]The variability within cluster and between clusters should be well incorporated in the design and analysis of a cluster trial. These degree of variability between clusters can be measured by the following appropriate measures: [2]

I. Coefficient of Variation, k

An alternative measure to the ICC is the coefficient of variation in the outcome, denoted by k. This is the ratio of the data's standard deviation to the mean (or the proportion or the rate) of the cluster outcomes. As the value of the standard deviation can be greater than the mean, values of k can exceed 1.[2]

II. Intracluster correlation coefficient, ρ

Intracluster correlation coeffiecient is commonly the preferred option to measure the between-cluster variability.

2.2.2 Intracluster correlation coefficient, ρ

In a cluster trial, one has to take into account the variance in the outcome within each cluster and also between cluster. The statistical measure of this intracluster dependence is the 'intracluster correlation coefficient' (ICC).[11, 31] The ICC is based on the relationship of the between to within-cluster variance and can be defined as the proportion of the total variation in the outcome that can be attributed to the difference between clusters.[7]

Generally, the value of the ICC can range from 0 to 1,[3, 31] An ICC of O means that individuals within the same cluster are no more similar than individuals from different clusters.[3, 7] This implies there is no cluster effect or in other words there is no between-cluster variability. An ICC of 1 would arise, when

all observations within a cluster are identical.[3, 7] This implies that individuals within the same clusters are correlated and there is no variation within clusters.[4, 17, 32]

Besides that, ICC is used to calculate the effective sample size for a cluster trial.[3] It is defined as the number of subjects required in an individually randomised trial to gain the same power as the cluster randomised trial.[3] The study by Campbell et al,2005 presented a formal analysis of factors that influence the magnitude of an ICC. The factors which influence the value of ICC in a research setting includes the type of variable, the study settings whether it is a primary or secondary care, the prevalence of the outcome and size of cluster. [11]

2.2.3 Design effect

In a cluster RCT, data is collected from a cluster sample of individuals in each treatment arm. Hence, cluster sampling provides less precise estimates of the outcome and less information as compared to simple random sampling. Design effect is used to measure the increase in variance resulting from the use of the cluster design.[17]

$$\boldsymbol{Design\ effect = \frac{Variance\ for\ cluster\ sampling}{Variance\ for\ simple\ random\ sampling}}$$

In short, design effect is the ratio of the total number of individuals required using cluster randomisation compared to the number required when using a conventional design.[3, 14, 32, 33]
Design effect can be represented by the equation below,

$$\boldsymbol{DE = 1 + (m - 1)\rho}$$

Where $\boldsymbol{\rho}$ is the Intracluster correlation coefficient and $\boldsymbol{m}$ the size of the cluster
The formula for the design effect takes the same form for both qualitative and binary outcomes.[17]

As the ICC value increases, the more important is the variation between cluster and therefore larger the design effect.[4, 32] Design effect increases with the value of ICC and the size of the cluster.[4] [3]The larger the design effect, larger the required sample size for the cluster trial to have the same power as a individually randomised study. [2, 4, 8, 14, 32]

2.3 Design Issues

2.3.1 Size of clusters

In a cluster study, a sample of individuals are selected from each cluster to measure the outcomes of interest. The size of the cluster depends on the statistical considerations and logistic or administrative problems in delivering the intervention.[17]

A study with a large number of clusters and fewer individuals within clusters will be able to distinguish intervention effects better compared to one that has fewer clusters but larger numbers of individuals within clusters.[4] This can be explained further by the formula for the design effect for a cluster randomised trial relative to an individually randomised trial as following:

$$DE = 1 + (m - 1)\,\rho$$

This formula shows that for a given total sample size, precision is maximum with a cluster size of m=1, where the design effect will be equivalent to an individual randomisation trial.[17] If ρ, the intracluster correlation coefficient, remains constant, the design effect increases with cluster size.[17] This implies that a large number of small clusters is statistically more efficient than a small numbers of large clusters.[4, 17] However, a large number of small clusters may not be effective, if there is a large amount of individual variation within the cluster.[4]

The estimate of the size of cluster plays a large role in a clustered study design. The size of cluster affects the calculation of the intracluster correlation (ICC), which later influences the design effect as well the final sample size calculation.

2.3.2 Unequal cluster sizes

In the real world, it is rather rare to encounter equal number of clusters in trials. This is due to natural variation in actual size of the clusters, variation in recruitment rate and loss of follow of subjects in a trial. The imbalance in cluster size reduces the power of the trial and has to be taken into account for in the sample size estimation.[34]

In imbalanced cluster sizes, estimates from the smaller clusters will be less precise and estimates from the larger clusters will be more precise.[35] The addition of individuals to larger clusters does not compensate for the loss of precision in smaller clusters.[35] Thus, as the cluster sizes become more unbalanced, the power of the study decreases.[35]

In studies with varying cluster sizes, cluster size is regarded as a random variable. When the cluster sizes are variable, the researcher may replace the value of *m* in the design effect with the use of average cluster size. However, this method underestimates the actual required sample size and increases the variation in the cluster sized.[36] To be safer, the largest expected cluster size in the sample is usually used. [36].

To account for variable cluster size, when the cluster size variability is large, Eldridge et al,2006 recommended that for situations where the accurate size of each cluster is not known, the value of the mean and standard deviation of the cluster size can be used to determine the sample size required.[35]This is done by inflating the design effect by multiplying the mean cluster size by $(cv^2 + 1)$, where cv is the coefficient of variation of cluster size.[35]

The value of the coefficient of variation of cluster size (cv), is defined as the ratio of the standard deviation of cluster size, S_m, to mean cluster size $\bar{m}$.[35]
The appropriate design effect for unequal clusters can then be rewritten as:[35]

$$DE = 1 + \{(cv^2 + 1)\bar{m} - 1\}\,\rho$$

This formula given by Eldridge et al (2006), may slightly overestimate the design effect, and works better when analyses are weighted by cluster size. More precise unequal cluster size calculations for various outcomes and situations is further described by Manatunga et al (2001), Jung et al (2003) and Pan (2001).[37]

Unequal cluster size rarely influences orthodontic research studies because the patient themselves are the unit of cluster. Lost of follow up of a patient results in lost of a whole unit of cluster.

2.3.3 Sample size considerations

There are a few issues which needs to be considered in the sample size calculations of a cluster study design. Firstly, the cluster design must be taken into account when estimating the sample size required.[33, 38]The standard sample size calculations are based on the assumption that the responses of individuals within clusters are independent and does not take the between cluster variation into account.[3] However, as mentioned above, in a cluster study design, the individuals or units within a group or cluster are not independent and individual within clusters tend to be more similar than individuals in different clusters.[4, 39] Therefore, the information provided by a given sample size in a cluster randomised trial is generally less compared to individually randomised trial. [40]As standard

sample size formulae do not account for all these factors, their direct use for cluster trials results in sample size estimates that are too small for a cluster trial. Understandably, there will be some loss of power due to randomisation by cluster rather than individual. [4] A cluster RCT with the same sample size as individual randomisation has a reduced power to detect an intervention effect, thus increasing the risk of resulting in a Type II error.[4]

Accurate sample size calculations for clustered designs require information relating to either the within-cluster correlations (ICC) or the between-cluster variability (coefficient of variation). Donner et al, proposed a simple method of sample size calculation for cluster trials. In order to achieve the required level of statistical power for a cluster trial, they proposed the inflation of the sample size of an individual randomisation trial by the design effect.[24 21]

To calculate the design effect, first estimate the intraclass correlation(ICC) followed by the estimation of design effect as explained above. After determining the design effect, the sample size of a cluster trial is calculated, where the number of participants required for individual randomisation is multiplied by the design effect.[4] The larger the ICC coefficient, the greater the design effect, and hence, a greater sample size will be required to match the power of a similar study by individuals.[4] This is the method of choice for randomised, two arm parallel-group design with fixed cluster sizes.

Kerry & Bland expressed the main difficulty in calculating sample size for cluster randomised studies, is obtaining an estimate of between cluster variation or intracluster correlations.[33] Estimation of variation between individuals can often be obtained from the literature. Unfortunately, even studies that use the cluster as the unit of analysis do not publish results in a way that the between-cluster variation can be estimated.[33] Recognising this problems, Donner recommended that authors should publish the cluster specific information and intracluster correlations, to enable other co-workers to use this information to plan further studies.[30]

2.3.4 Ethical and consent considerations

As a cluster RCT involves a larger number of sample size and done at a level involving many groups, it can be rather challenging to provide individual choices for interventions that are implemented.[34] Hence, cluster RCTs present difficulty in regards to representations and ethical considerations (Medical Research Council, 2002).

The ethical concerns and challenges encountered in obtaining informed consent in cluster randomised trails has been well explored by Edward et al, 1999 where a comparison between the individual cluster

trials and cluster cluster trials showed the likelihood of obtaining informed consent is linked to the level at which study interventions are administered.[40] Cluster trials are done to be able to study group affects, logistic demands and to prevent contamination.[32] Study interventions in individual cluster trials are directed and studied at the individual cluster members, very similar to an individually randomised trial.[41] Hence, individuals within clusters can provide consent individually for the treatment offered within the cluster and informed consent should be obtained in individual-cluster trials.[41]

The study intervention in a cluster trial is applied to an entire cluster, making it impossible for an individual member to not participate in an intervention. Hence, this makes it difficult to obtain informed consent from an individual member in a cluster. In a cluster cluster trial, "individuals cannot act independently and the autonomy principal is lost".[40, 41] Furthermore, informed consent for randomisation in CRTs is difficult to acquire because individual cluster members could be randomised even before they are identified. Also, the large sample size of individual members in a cluster adds to the difficulty in obtaining consent. In summary, there are many factors that impede the ability to acquire informed consent from all study participants in CRT's involving cluster level intervention and this further contributes to the ethical consideration to CRTs.[41]

A solution to this ethical challenge involves a provision of waiver of consent found in The Council of *International Organization of Medical Science International Ethical guidelines for Biomedical Research Involving Human Subjects* – a widely acknowledged commentary on the Declaration of Helsinki- contains the following provision. ' when the research design involves no more than minimal risk and a requirement of individual informed consent would make the conduct of the research impracticable, the ethical review committee may waive some or all of the elements of informed consent'.[42] This provision allows a research committee to waiver consent when the individual risk is minimal and it is not possible to obtain consent from each individual in the study.[41]

With this, research committees are allowed to carry out procedures which does not include or alters some aspects of informed consent or even waives the requirements to obtain informed consent, provided the research committee meets the following criteria:[41]

a) 'The research involves no more than minimal risk to the subjects' [41]
b) 'The waiver/alterations will not adversely affect the rights and welfare of the subjects
   By this it means, the research ethics committee must ensure that the welfare of the subjects, are not adversely affected by the waiver of consent'.[41]
c) 'The research could not be practicably carried out without the waver or alteration'[41]

d) 'Whenever appropriate, the subjects will be provided with additional pertinent information. Although a waiver of consent is granted on grounds that obtaining informed consent is not possible, effort should be taken to inform cluster members of the existence of the study whenever feasible.'[41]

Overall, informed consent of research subjects is required to comply with the ethical principle of respect for person. However, there are certain challenges to do so due to certain aspects of cluster study design including cluster sample size, randomisation and cluster level intervention.[41] McRae et al, 2011 has addressed these challenges by using a waiver of consent and ensuring the conditions are kept. Furthermore, if it is not possible to approach individual subject at the time of randomisation, consent for randomisation may not be necessary.[32] Additionally, adequate information of an intervention in the trial arm needs to be provided to each individual after cluster randomisation. However, it is not necessary to provide individuals information about other trial arm interventions that does not involve them.[42] Also, a passive consent is not a valid informed consent. Lastly, it is still necessary to obtain informed consent when health professionals participate as subjects in research.[41] In summary, ethical issues in cluster RCTs need to be addressed appropriately. However, in the majority of the cases where clustering is observed in orthodontic research, it involves clusters of teeth in an individual's mouth. Hence, it is rarely an ethical issue.[3]

2.4 Analytical methods

2.4.1 P-value

The use of probability levels (P values) and statistical significance in testing the interpretation of research findings is among the universal tools of scientific practice. P-value as an idea of significance testing was introduced by R. A. Fisher in his seminal work the 'Statistical Methods for Research Workers'43. The concept of p-values as a measure to be employed in scientific practice was further illustrated by Fisher in the 'The Designs of Experiments' where the 'lady tasting tea' experiment was presented.[44] In that particular experiment, a subject that claimed to be able to distinguish between two different variations of tea and milk being mixed together in its preparation when consuming the cup of tea. 8 cups were presented, with 4 each being prepared by a single variation of tea and milk being mixed, in which the subject, who was blindfolded, would have to segregate them accordingly based on their variation of mixing. Given 8 cups of tea, there existed 70 different possible combinations with only 1 right combination, giving us a probability of 1/ 70 . Thus, Fisher suggested that if such a result were obtained, it would be 'surprising' enough given what we initially assumed,

that it is not possible to distinguish between the two different variations, to warrant further investigation into the relationship between the variables.

In simple terms, the p-value is defined as the probability of the observed results, plus more extreme results (either greater than or less than) of a particular random variable, based on the initial assumption that a null hypothesis, determined at the beginning of the experiment, is indeed true.[45] Thus, the p-value serves as an index that measures the significance of a relationship between the control variable and the responding variable in an experiment, based on the assumption that the null hypothesis is true.[46]

The threshold value level of significance, denoted by alpha (α), is arbitrary and is usually set in advance.[47] It takes the value between 0 and 1. A low p-value (closer to 0) suggests the probability of obtaining the observed difference is low, given the null-hypothesis is actually 'true'. In other words, the smaller the P value, the stronger the evidence against the null hypothesis actually being a valid relationship, based on the evidence that had been collected from a particular experiment. Value close to 1 indicates there is no difference between the groups.
In scientific papers, the level of significance that is taken to be significant is conventionally set at $P<0.05$.[45] This is equivalent to a chance of 5 in 100 or 1 in 20 that such a result could have occurred by chance alone under the assumption that the null hypothesis is true. Therefore, if the p value is less than 0.05, there is sufficient reason to reject the null hypothesis based on the evidence available as the likelihood of observing the results, in the 'true' null hypothesis, is appreciably low. We then reject the null hypothesis and conclude that the results are significant at the 5% level.

In contrast, if the p-value is equal or greater than 0.05, it is concluded as there being insufficient evidence to reject the null hypothesis, hence, the results are not significant at the 5% level. However, such a result does not conclude that the null hypothesis is true. It simply concludes that, based on our 'limited' sample, the findings of our evidence are not strong enough to suggest that the relationship suggested by the null hypothesis can indeed be rejected.[48]

Neyman and Pearson disliked this approach by Fisher that was deemed subjective in nature going against the objectivity demanded by the scientific method in proving or disproving a particular hypothesis. They suggested that two types of errors could exist when interpreting results.[42]
The risk of experiencing a type I and type II error is always present, due to the nature of how hypothesis tests are carried out. The type I error is also known as false positive error in hypothesis

testing. It is defined as rejecting the null hypothesis when it is in fact true and concluding that an effect exists where there indeed isn't.[42,49] In cases such as this, the researcher would have wrongly rejected the null hypothesis, even as the findings of their evidence and a set threshold level of likelihood that had been initially set suggest so. It is equivalent to the threshold used for statistical significance, which conventionally stands at 0.05, which is again represented by alpha (α).

The type II error is also known as the false negative in hypothesis testing. The researcher does not reject the null hypothesis when it is false and concludes that there is no effect when there is a true effect exist. [42,49-51]The probability of making a type II error is denoted by Beta (β). Its counterpart with the equation 1-β is the power of the test. The power is the measure of the possibility of detecting possible difference between groups provided that such a difference exist. Normally, β is arbitrarily set at 0.1 or 0.2, which means a study has either 90% or 80% power to detect a given difference at a specific degree of significance. This is similar to power calculation done in a trial, to ensure that the study is large enough to allow both type I and type II error rates to be small.[45]

2.4.2 Limitation of p-value

Obtaining 'significant' or 'non-significant' results according to p-value should not be the ultimate aim of performing statistical analyses.[42] Kee-Seng Chia described an obsession with P-value in his article 'Significant-it is'. This practise of over-dependant on the p-values when interpreting results in the dichotomy of significant or non-significant leads to erroneous conclusions.[50]

The p value itself is influenced by the sample size and the variances. The p value becomes smaller when there is a larger sample size and a smaller standard deviation. However, a small p value does not necessarily indicate a large intervention effect and a larger p value does not advocate a lack of effect.[47,52] A small differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects may be statistically non-significant due to a small sample size.[47,52] The p value therefore provides no insight into practical relevance due to the lack of the effect size, range and the clinical effectiveness of the observed results.[52,54,55]

In short, a statistically significant result could possibly be clinically irrelevant and vice versa.[50]
Due to the limitation on p-values, most authors have advocated the use of confidence interval on the resulting p-values.[42,50,52]

2.4.3 Confidence intervals

Confidence intervals provide us with an upper and lower limit around the parameter looked at in a study. It gives information on the effect point estimate between the study groups which then helps to determine whether the observed differences are suggestive of true benefits of the treatment. The effect point estimate is usually set at 95 percent, in which we are 95 percent confident that the true population effect lies between these two points. This valuable information on the magnitude of the differences between the study groups aids in making clinical decisions.[56]

The width of the CI quantifies the precision of the results.[56] It is influenced by its standard error, which in turn depends on the standard deviation and sample size.[53,57] Small sample size leads to wide confidence intervals with less precision.[55] Increased sample size narrows the width of the CIs around the same size of effect, thus increasing precision.[53,56,57] This is opposite to p-value, where increasing the sample size lowers the P value.[47,56]

Polychronopoulou et al conducted a search on the orthodontic literature, aiming to determine the frequency of the reporting of confidence intervals in orthodontic journals.[56] It was rather disappointing to know only 6% of the included articles reported on the confidence interval along with P-value.[56]

2.4.4 Influence of clustering on p-value and confidence interval

As discussed above, the key difference between a cluster trial and an individually randomised trial is that groups of individuals/subjects are allocated to the treatment arms. The interventions are randomly allocated to these groups/clusters instead of individual/subject level.[17] Therefore, the cluster constitutes the experimental unit in a cluster trial.[17]

As the observations in the same cluster are likely to be correlated, the analysis of a cluster trials must take into account of the clustered nature of the data. Treating the individual subject/teeth as independent and discounting for the correlated nature of the data increases the chance of getting significant results, which are false.[5 59]This is because in a cluster trial, the size of the standard errors increases thus widening the confidence interval and increasing ***P*** values compared to a conventional trial of the same size there by reducing the power as the effective sample size is reduced.[2, 3, 10, 29]

This can be explained better with the basic form of statistical test formula.[59]

Test statistic = estimate / standard error (d/se)

Where se= sd/$\sqrt{n}$, *d= estimate*, *sd* is the standard deviation and $\sqrt{n}$ the square root on ***n*** (sample size). It is worth to note standard error is directly related to the variability of the observations and inversely

related to the sample size. From the formula, as the sample size increases the value of the test statistic increases as well.[1] This follows by a lower ***P*** value, increasing the chance of observing a statistically significant result, which is a Type 1 error where researcher rejects the null hypothesis when it is true and concludes that an effect exists when it does not.[1] In short, the p-value becomes smaller when there is a larger sample size and a smaller standard deviation.

Furthermore, in a clustered design, the amount of the information contributed by each cluster is reversely proportional to the within cluster correlation of the observations.[1, 30] The larger the correlation of the within clusters observations, the lower the contribution of each individual/subject to the analysis.[1] As the contribution of each individual/subject decreases, so does the effective sample size. Base on the statistical formula, as the sample size decreases, standard error increases resulting in smaller test statistic and larger *P* values.[2] In short, correlated data when treated as uncorrelated(no clustering present in the data) gives significant results and when correctly treated as correlated gives non-significant results.[1, 2]

Clustering also influences the width of confidence intervals. The larger the sample size, the narrower and more precise is the confidence interval. A cluster study with similar effective sample size to conventional study presents with wider confidence interval. Similar to p-value, correlated data treated as uncorrelated gives a narrower confidence interval resulting in incorrect inferences.

2.4.5 Multiple Hypothesis Testing

In clinical studies, researchers may wish to compare groups on multiple different outcomes and therefore perform multiple statistical test. However, as the number of significance tests on a data set increases, there is a greater possibility of a false positive result.[60] Every statistical test comes with an inherent false positive, or type I error rate—which is equal to the threshold set for statistical significance, generally 0.05. However, this is the error rate for one test only.[62] When more than one test is run, the overall type I error rate is much greater than 5%.[61]

Multiple hypothesis testing is commonly performed in studies comparing multiple outcomes, multiple predictors, repeated measures over a period of time on the same outcome, subgroup analysis and interim analysis of treatment effect at different stages of treatment.[61] Studies involving repeated measurements done on the same subject are also subjected to clustering effects, as the measurements belong to the same subject.

To prevent the occurrence of Type I error in studies with multiple hypothesis tests, a statistical correction should be performed to account for the number of statistical tests run. Several correction methods exist such as Bonferroni, Sidak, Benjamini & Hochberg and Holm's for specified multiple hypothesis testing.The most simple and popular method used by researchers is Bonferroni correction. The basic idea is to preserve the overall type I error rate at .05 by lowering the threshold for statistical significance to lower than 0.05.[61]

On the other hand, applying correction for multiple hypothesis testing to reduce type I error can result in studies with reduced statistical power which means that there is a reduced probability of rejecting the null hypothesis [H0] given that null hypothesis is false (type II error). In other words, it reduces the likelihood that the tests will identify the true differences between the groups.[51]

2.4.6 Data analysis consideration for cluster RCT

As the cluster is the experimental unit in a cluster trial, observation can be made at different levels and thus there can be several different types of observational unit. There are two main approaches to the analysis of cluster trials, involving two treatment arms with no matching or stratification.[2,8,17,63]

i. Analysis at cluster level
ii. Analysis at Unit/Subject level

2.4.6.1 Cluster level analysis

In a cluster level analyses, the cluster is the unit of analysis.[1, 2, 8] This can be in terms of the mean of the outcome or a proportion.[31] As each cluster then provides only one data point, the data can then be considered to be independent and standard statistical tests can be used.[31]This is a two-stage process.[17] Firstly, a summary measure is obtained for each cluster, which is usually based on data collected on outcome among subjects in that cluster. This is followed by simple statistical tests on the cluster-specific measures to compare the effect of estimate between the treatment arms.[2, 8, 17]

There are advantages and disadvantages to this level of analysis. Because the clusters are the experimental units in a cluster trials, it is a logical to obtain a measure of the total outcome for each of this unit and then compare the means between treatment arms.[11] Furthermore, this approach has been shown to be robust, as it can be applied to any outcome variable and allows the construction of a statistical inference.[13]

However, as the analyses is based on cluster summaries, there is some loss of information when the data are reduced to a set of summary measures.[64] For example, in a study with bracket failure as the outcome measure, the rate of bracket failure may differ between maxillary and mandibular teeth or between anterior and posterior teeth.[65] These differences are however not reflected on the number of failed brackets per patient, thus, there is indeed some information loss. In a cluster level analyses, the effects of the intervention and of other covariates are not analysed together in the same regression model. Instead, it is a two-stage approach in which the cluster level summaries at the second stage have already had the effects of other covariates removed.[11] As advocated by Hayes and Moultan (2017), several statistical methods can be employed in a cluster trial such as two-sample t-test, weighted t-test, Wilcoxon's rank sum test or a permutation test. The choice of test will depend on the nature of the research question and the distribution of the cluster summary data.

2.4.6.2 Unit level analysis

In unit level analysis, analysis is carried out on the observations within a cluster.

Analyses using unit of analyses is commonly done using a number of regression models that adjust for the clustering effect.[1, 2] This single-stage method allows to analyse the effects of intervention and other covariates in the same model. Here, all inherent correlation within clusters are modelled explicitly, allowing a 'correct' model to be obtained. This helps to increase the statistical power of the analysis.[31]

The main advantage of this individual level analysis is the effects of modelled covariates can be estimated simultaneously with the intervention effects in the same regression model.[11] In contrast to cluster analysis, where the comparisons of cluster-level summary is done in the second stage after the effects of other covariates have been removed from the first model. In short, it allows more direct examination of the joint effects of cluster-level and individual-level predictors.[20] Individual level analysis allows to look for individual results while accounting for clustering effects, preventing loss of individual data. On the downside, individual level analyses are not accurate for small number of clusters.[11]

Hayes and Moultan (2009), have suggested various regression methods for individual level analysis including random effect models, generalised estimating equations and repeated measures of ANOVA. Repeated measures ANOVAs are appropriate when clustering is based on timepoints, and complete data is available. Random effect models and generalised estimating equations are more flexible in allowing for missing data and different types of clustering, such as clustering by tooth.

2.5 Reporting and Interpretation

2.5.1. Reporting of cluster randomised trials

In the past, trial reports did not always meet the highest standards. Hence, the editors of leading medical journals sponsored the publication of the Standards of Reporting of Trial (CONSORT) statement in 1996. The aim was to improve the reporting of randomised controlled trials. The original CONSORT guidelines were designed for use with individually randomised trials, and did not consider the features of a cluster randomised trials that need to be addressed when reporting such trials. Subsequently, an extended statement was published that provided guidelines on the reporting for cluster randomised trials.[66]

In 2008, the CONSORT group produced a separate reporting checklist for abstracts of randomised controlled trials, which presented a minimum list of essential items that should be reported within a trial abstract. Later, in 2010, an updated and extended CONSORT statement was published to integrate the important advances in the methodology for cluster trials since 2004. The updated CONSORT 2010 statement includes a checklist of 25 items, which should be included in the trial report. Most journals now require all reports of cluster randomised trials, conform to the guidelines in the Consolidated Standards of Reporting Trials (CONSORT).[17]

The Consolidated Standards Of Reporting Trials statement extension to cluster RCTs requires reports of cluster trial should include the following additional information:[3,66]

a) The title of the trial should clearly identify as a cluster randomised trial.[66]
b) In the abstract, the design of a cluster study should be clear, specifying that allocation was based on cluster. It should include information on the method of randomisation, number of clusters and the level of analysis of the primary outcome.[66]
c) The rationale for adopting a cluster design should be outlined in the background.[66]
d) The description of the specific objectives and hypothesis should describe whether they pertain to the individual/subject level, cluster level or both. If objective or hypothesis are targeted at cluster level, analysis and interpretation of results should follow at cluster level as well.[66]
e) The trial design should include the general description of the trial and descriptions of how the design features are applied to clusters. Whether the cluster randomised design is parallel, matched pair, or other and whether the treatments have a implications for the appropriate analysis of the outcome data.[66]

f) Two sets of eligibility criteria should be reported. This includes the eligibility of the clusters to be included in the trial and the eligibility of individual subject to be included in the clusters.[66]

g) Description on whether the intervention was targeted at the cluster level or the individual/subject level and level at which outcome is measured. The level of intervention influences the analysis of outcome data. Therefore, it is important for the trial to be explicit about the level at which outcomes are measured.[66]

h) How the effects of clustering were incorporated into the sample size calculation. Detailed information on the method of calculation, size of cluster, number of clusters, and value of intracluster correlation coefficient should be reported. In contrast to individually randomised trials, sample size calculations in a cluster trial need to take account of the between-cluster variability.[66]

i) Steps of the random allocation process from generation to implementation should be reported adequately. The randomisation process in a cluster trial should include inclusion and allocation of clusters as well as the inclusion of cluster members. Therefore, the implementation process adopted for each step need to be outlined separately and information on the mechanism by which individual participants were included in clusters for the purposes of the trial.[66]

j) Details from whom consent was sought and whether consent was sought before or after randomisation should be reported. The level of consent highly depends on the level of intervention administered. The level of consent sought and issues around informed consent in cluster randomised trials have been described above.[66]

k) How the effects of clustering were incorporated into the data analysis.
Statistical methods used to compare groups for primary outcomes should indicate how clustering was taken into account and methods for additional analyses done. As discussed above, a wider range of statistical methods can be applied to cluster randomised trials compared to individually randomised trials.[66]

l) A flow diagram of the clusters and number of individual subjects throughout the trial should be reported at each stage. Specifically, for each group report on the number of clusters and participants randomly assigned, receiving intended treatment, completing the study and analysed for the primary outcome should be included. A CONSORT flow diagram with clustered data can be presented based only on clusters, only on individual participants or on both.[66]

m) When reporting the results of a cluster randomised trial, point estimates with confidence intervals should be reported for primary outcomes at cluster or individual level as applicable.[67] Additionally, specify the assumptions used when estimating the size of cluster and within-cluster samples in the trial. All this information provided, together with cluster size and design

effect allows readers to access the appropriateness of the sample size calculations. Sample size calculations of a cluster randomised trial, requires estimates of ICC .Obtaining the value of ICC has been recognised as the main difficulty in calculating the sample size for a cluster trial.[32] This is because, most studies do not publish the value of ICC or the assumptions in estimating the variation between individual/subjects.[32] Therefore, the extended CONSORT statement guidelines recommend that observed values of ICC in a cluster trial should be reported as well. This would enable researchers to accumulate evidence on appropriate ICC values in planning future cluster studies.

2.6 Interpretation of cluster randomised trials

The interpretation of the results from cluster randomised trials vary from the individually randomised trials as the conclusions are related to the clusters, subjects in those clusters or to both. Failure to account for clustering can lead to inaccurate results and potentially misleading conclusions especially if the interpretation is based solely on *P* values.[1]

In summary, the effect of clustering has to be taken into account in the design, conduct, analysis, reporting and interpretation of cluster trial. When planning a cluster trial, the main issues such as sample size requirement, size of each cluster, blinding, allocation concealment, and level of consent, method of data analysis should be addressed from the very beginning. However, a previous study by Koletsi et al(2012) has reported a large number of orthodontic literature presenting with clustering effects but did not account for these effects in data analysis, resulting in misleading conclusions.

Chapter 3: Study aim and objective

3.1 Study aim

The aim of this study is:

- To examine the extent of clustering effects in the recent published orthodontic literature and to determine the frequency by which clustered designs are correctly addressed in the statistical analysis.

3.2 Study Objectives:

The objectives of this study were to:

- To search the orthodontic literature for studies presenting with clustering effects.
- Quantify studies presenting with clustering effects in the orthodontic literature.
- To determine the frequency by which clustered design articles, accounted for the clustering effects during statistical analysis.
- To present narrative and tabular summaries of the number of articles considered for clustering, number of articles presented with clustering effects, number of articles which accounted for the clustering effects as well as total number of articles which did not account for the clustering effects in the statistical analysis.
- Describe statistical methods used to account for the clustering effects in the statistical analysis
- To determine the potential association of the study characteristics such as journal of publication, continent of origin, type of study, number of authors, collaboration with a statistician, single or multicentre study, statistical significance of the results with appropriate management of the clustering effects in statistical analysis.

Chapter 4: Methodological Framework

4.1 Study design

This was a retrospective, observational study looking at a sample of published orthodontic articles in three orthodontic journals over a two- year period from 1st January 2016 to 31st December 2017. Figure I summarises the methods employed in this study.

4.2 Study selection criteria:

Inclusion Criteria:

Three major orthodontic journals were included in this study. This included the American Journals of Orthodontics and Dentofacial Orthopedics (AJO-DO) (formerly known as American Journal of Orthodontics), Angle Orthodontist (AO) and European Journal of Orthodontics (EJO). The rational for selecting only these three journals was to use the similar sample which were used in the study done by Koletsi et al to be able to assess if there assess if there is a change in the proportion of orthodontic studies which account for clustering in statistical analysis. All articles published in these journals in the year of 2016 and 2017 issues were eligible for inclusion in this study.

Hence, the following issues were included in this study.

i. Volume 150 and 149 for the year of 2016, AJO-DO
ii. Volume 151 and 152 for the year of 2017, AJO-DO
iii. Volume 86 for the year of 2016 , AO
iv. Volume 87 for the year of 2017, AO
v. Volume 38 for the year of 2016, EJO
vi. Volume 39 for the year of 2017, EJO

In total, the content of 48 issues were assessed including 24 from the AJO-DO, 12 from AO and 12 from EJO.

Exclusion Criteria:

The following articles were excluded:

i. Studies involving animals
ii. *In vitro* studies
iii. Articles evaluating technique descriptions
iv. Studies not involving patients, such as simulation studies
v. Case reports

vi. Case series
vii. Review articles
viii. Letters to editors, book chapters, abstracts and commentaries
ix. duplicate studies (studies originating from the same subjects by the same investigators but published in different journals)

4.3 Search methods for identification of studies:

Hand searching

The full text of the articles and content of the above mentioned journals published in 2016 and 2017 were hand searched systematically by the first author (BB) in order to identify for published articles in which clustering effects were evident from the methodology report. Using library resources, all issues of AJODO, EJO and AO were accessible. They were accessed as electric journals via the University of Liverpool library account. Hence, an online search for each of the journal issue on the respective websites of the included journals were carried out by the first author. (BB) Print out of the issue synopsis were used for identification of papers.

Language

All articles from the included three journals were in English. Therefore, no effort for translating non-English papers was required.

4.4 Pilot study

Prior to the commencement of the article search, BB discussed with the research supervisors (GB,NF) on the article selection and data to be extracted from the included studies.

Search on few journal issues were carried out by first author (BB) along with research supervisor(GB) during research meetings. The pilot study included one issue of publication from each journal. This allowed the first author (BB) to learn to identify articles presenting with clustering effects and gave an exposure on interpretation of the statistical analysis. Furthermore, this allowed to identify any potential problems in the study design and gave an exposure of the variability of articles presenting with clustering effects. This was done until good level of understanding in extracting articles presenting with clustering effects was obtained by the first author (BB). The data extraction forms were finalised through discussion with supervisors during the pilot study and were subsequently used in the present study.

4.5 Selection process

The selection process of the relevant articles in the above mentioned journals involved multiple stages. After initial piloting, the first author (BB) independently assessed full text articles published in these selected journals against the inclusion and exclusion criteria to identify potentially relevant research publication. Editorials, reviews, and case reports could be identified from the title or abstract and later excluded. The methodology of each article was assessed to identify publication in which clustering effects existed in the study. When an articles was deemed to present with clustering effects, the results and the method of statistical analysis was explored further in detail to identify articles which have accounted for these effects in the statistical analysis. A maximum of two issues of journals were assessed at any one time with a view to prevent errors due to fatigue.

All the articles were later discussed with the research supervisor (GB) during research meeting for confirmation to be included in the study and further assessment on the statistical analysis. Disagreements were resolved by thorough review of the article and further discussion between BB and GB. We consulted a third review author (NF) if we could not resolve disagreements.

The number of articles considered to have clustering effects, articles which accounted for the clustering and articles which did not account for the clustering in the statistical analysis were documented in a tabular summary form (Appendix 2).
If an article was deemed suitable and presented with clustering effects, it was further assessed to note on the following parameters.

i. Journal of publication

   The articles were classified according to their Journal of publication of either, the American Journals of Orthodontics and Dentofacial Orthodontics (AJODO), European Journal of Orthodontics (EJO) and Angle Orthodontist (AO).

ii. Type of study

   Articles were categorised as interventional or observational study base on method the study was carried out. Interventional study includes any human trial (clinical or randomised clinical trial) involving an experiment or other interventions with a control group. It is often a prospective study which is specifically tailored to evaluate direct impacts of intervention, of a treatment or preventive measure. Articles were classified as observational for any ecological design, case control, cohort and cross-sectional study, either prospective or retrospective.

iii. Region of authorship/ Geographical Area

Articles were categorised according to the geographic region of the first author. If the published studies had authors from more than one country, only the country of origin of the first author was recorded. The continent of the authorship was subdivided into the following:

1. America
2. Europe
3. Asia
4. Other

Articles from North and South America were kept together in the continent 'America' category. Articles from countries which did not belong to either the America, Europe or Asia continent were categorised in the 'Other' category.

iv. Single or multicenter study

Study of single or multicentre study was recorded. Studies conducted at only one site or hospital or medical centre, were recorded as single centre. Studies conducted using a single protocol, at two or more sites, each with its own clinical investigator was recorded as a multicentre study. This was assessed from the affiliation details and any other information provided in the methodology section, on where the study was conducted, and data was collected.

v. Number of authors in the publication

This was assessed from the affiliation details provided at the start of the article. It was categorised to less than three, four or more than five researchers in the study.

vi. Involvement of statistician

Collaboration with statistician was determined by the affiliated information given by the authors in the article. When there was no information of an involvement of a statistician in the article, a google search on the names of the associated authors was done to note on the involvement of a statistician.

vii. Statistical significance

Also, statistical significance of the primary outcome of the study was noted. Statistical significance is the likelihood that a research results is true and not merely a matter of chance. $P < 0.05$ was considered as statistically significant, unless noted otherwise. The results were compared to the p-value mentioned in the articles to detect a clinically significant difference and evidence to reject the null hypothesis. This was a binary column of yes or no.

viii. Statistical method used

The statistical method used in the articles presenting with clustering effects was assessed and categorized to the following categories:

a) Analysis of variance (ANOVA) category includes k-way ANOVA, multiple analysis of variance and non-parametric ANOVA
b) Chi-square category includes chi-square, Fisher's exact test, Homogeneity test and Mc Nemar's test
c) Mixed models category includes mixed models, Friedman/repeated measures ANOVA and Generalised estimating equations.
d) T-test category includes independent and paired t-test, non-parametric equivalents such as Mann-Whitney, Wilcoxon and Signed rank tests
e) Survival category includes Cox regression, Kaplan-Meier and Log rank tests
f) No statistics category includes descriptive statistics or nothing reported
g) Correlations
h) Linear regression
i) Logistic regression

Along with the above mentioned eight parameters, the following two additional parameters were recorded as well:

i. Sample size calculation

Articles presenting with clustering effects were accessed if sample size calculation was reported. The sample size used in a study is determine based on the expense of data collection, and the need to have sufficient statistical power.

This was a binary column of yes or no. It will be considered yes for articles which report on the number of subjects required to achieve the targeted statistical power and significance. Articles which accounted for the clustering effects in the sample size calculation were noted in the remark column.

ii. Cluster type

The studies were classified according to the type of clusters presented in the study. This included either multiple teeth, multiple time points, multiple assessors and others. This column was to identify the common cluster type adopted in most studies presenting with clustering effects.

4.6 Data extraction and items

A pre-designed and piloted data collection sheet was prepared to extract relevant data from each included study. This allowed systematic data collection from each individual study and to record the additional parameters of the studies with clustering effects. All data collected was saved electronically.

A structured table in a Word format (Appendix II) was prepared by the first author (BB) to record the following information:

i. The journal and volume of publication
ii. The issue and month of publication
iii. The total number of articles identified in the issue or the month of publication
iv. Number of articles excluded in the issue or month of publication
v. Number of articles considered and included in the study
vi. Number of articles considered to have clustering effects base on the methodology reported
vii. Number of articles which presented with clustering effects and accounted for those effects during data analysis
viii. Number of articles which presented with clustering effects and did not account for those effects during data analysis

Besides that, a structured data extraction form (Appendix III) was used to systematically collect the information of the additional parameter from the articles considered to have clustering effects. Each of the articled was assessed on the following items:

i. journal of publication
ii. title of article
iii. continent of origin
iv. Involvement of a statistician
v. type of study
vi. sample size calculation
vii. Cluster type
viii. statistical significance of the results
ix. statistical method used
x. Additional column of remarks will be included to allow space for comments.

All outcome data was extracted and recorded. In addition, input from research supervisors (GB) was obtained, if there was any uncertainty during the data extraction stage.

4.7 Assessment of reliability

After two months into data collection, 10 percent of the total number of articles were reassessed by the first author to determine the intra-rater reliability. This included a random pick of an issue of publication from AO and EJO and two issues from AJODO. The intra-rater reliability tests were tabulated and assessed using kappa statistics and percentage agreement.

The following four issues were assessed:

I. Volume 151, Number 2, February 2017 from AJODO
II. Volume 152, Number 1, July 2017 from AJODO
III. Volume 87, Issue 2, March 2017 from AO
IV. Volume 39, Number 2, April 2017 from EJO

Inter-rater reliability assessment was not done because all the articles included in the final analysis were discussed by GB. If the level of agreement was low between the examiners (BB, GB), further discussion was arranged.

4.8 Data entry

The data extracted was entered in two documents.
A structured table in word format was used to summarise the numbers of articles identified, articles excluded and articles presenting with clustering effects, from each journal issue. Articles with clustering in the study design were further divided to number of articles which have accounted for clustering, not accounted for clustering in the statistical analysis and articles analysed each of the outcome separately. (Appendix 1)

A customised Microsoft Excel spreadsheet (Version 15.14, Year 2015, Microsoft, Microsoft Office 2015, Microsoft Corporation, Redmond, USA) was used to systematically collect the information of the additional parameter from the articles considered to have clustering effects. (Appendix 2)

4.9 Quality assessment

During the stage of data collection, there were no attempts made to access the quality of the individual articles from the study sample. This was considered to be out of the remit of the aim and objectives of the research to make further evaluation of this aspect.

4.10 Statistical methods

Descriptive statistics was used to analyse the characteristics of the articles presenting with clustering effects. Values were presented in raw data and percentages. A tabular summary of the frequencies of statistical methods used in the included articles and articles which correctly accounted for the clustering effects in the statistical analysis were presented.

Multivariable and univariable logistic regression analyses were undertaken to determine the association between the clustering effects (dependent variable) and the independent variables. This included the journal of publication, continent of origin, type of study, number of authors, collaboration with a statistician, single or multicentre study, sample size reporting and statistical significance of the results.
Statistical significance was set at 0.05. Backward elimination was applied to access variables that were associated with the outcome.

4.11 Statistical analysis

This was undertaken by using IBM SPSS statistics, Version 25.0 (Armonk, NY:IBM Corp)

4.12 Ethical Implication

This was a retrospective observational study using the raw data from previously published orthodontic literature. Since there was no contact with study subjects and no patient identifiable data used, ethical consideration was considered to be unnecessary.

Chapter 5: Results

The results are presented in the following seven sections:

5.1 Results of the search

5.2 Results of articles accounting for the clustering effects in the statistical analysis

5.3 Characteristics of the included articles and factors influencing accounting of clustering in statistical analysis

5.4 Summary of statistical methods used in articles presenting with clustering effects in the study design

5.5 Univariable and multivariable logistic regression for articles accounting versus non-accounting for clustering effects when including 'separate analyses' articles

5.6 Univariable and multivariable logistic regression for articles accounting versus non-accounting for clustering effects in statistical analysis

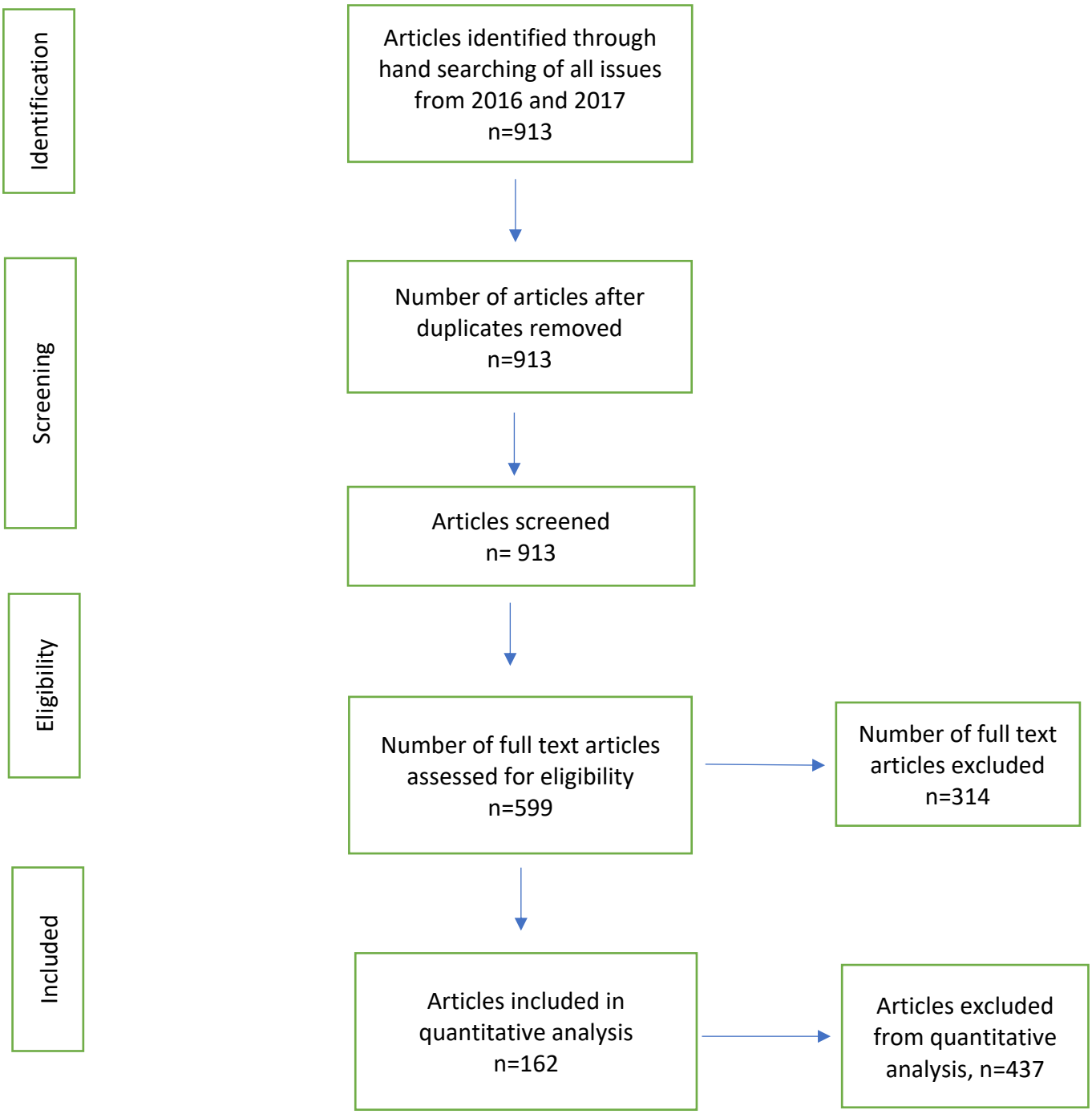
5.7 Effects of clustering on finding significant results

5.8 Inter and Intra reliability assessment

5.1 Results of the search

The flowchart indicating the search results is shown in Figure 5.1

Figure 5.1: Flowchart indicating the search results



5.1.1 Overall number of the articles identified

All issues of the AJODO, AO and EJO published in 2016 and 2017 were hand searched. A total number of 913 articles were identified from 48 issues of journals, as illustrated in Table 5.1. Of these, 478 articles were from the AJODO journal, which makes up 52.4% of the total number of identified articles. Most articles identified are from AJO-DO because it is a journal which is published monthly. On the other hand, AO and EJO are published bimonthly. 251 (27.5%) articles were identified from Angle Orthodontist and remaining 184 (20.2%) of the articles were identified from EJO. Of the identified 913 articles, 470 (51.5%) articles were published in 2016 and 443(48.5%) were published in 2017.

Table 5.1: Overall number of articles identified from AJODO, AO, EJO journals

Journal	Number of issues	Number of articles in 2016	Number of articles in 2017	Total number of articles
AJODO	24	239	239	478 (52.4%)
AO	12	137	114	251 (27.5%)
EJO	12	94	90	184 (20.2%)
AJODO+AO+EJO	48	470 (51.5%)	443 (48.5%)	913

5.1.2 Overall number of articles fulfilling the eligibility criteria

After applying the pre-defined exclusion criteria, 314 articles were excluded. The pre-determined inclusion and exclusion criteria were used to screen the full-text articles. Following that, there was a total 599 articles that fulfilled the eligibility criteria, as illustrated in Table 5.2. Of these, 253(42.2%) articles were published in AJODO, 205 (34.2%) articles published in AO and 141 (23.5%) articles were from EJO. Of the 599 articles eligible at this stage of data collection, 313 (52.2%) articles were published in 2016 and 286 (47.8%) were published in 2017.

Table 5.2: Overall number of articles fulfilling the eligibility criteria

Journal	Number of issues	Number of articles in 2016	Number of articles in 2017	Total number of articles
AJODO	24	129	124	253 (42.2%)
AO	12	112	93	205 (34.2%)
EJO	12	72	69	141 (23.5%)
AJODO+AO+EJO	48	313 (52.2%)	286 (47.8%)	599

5.1.3: Number of articles associated with clustering which were included in the final analysis

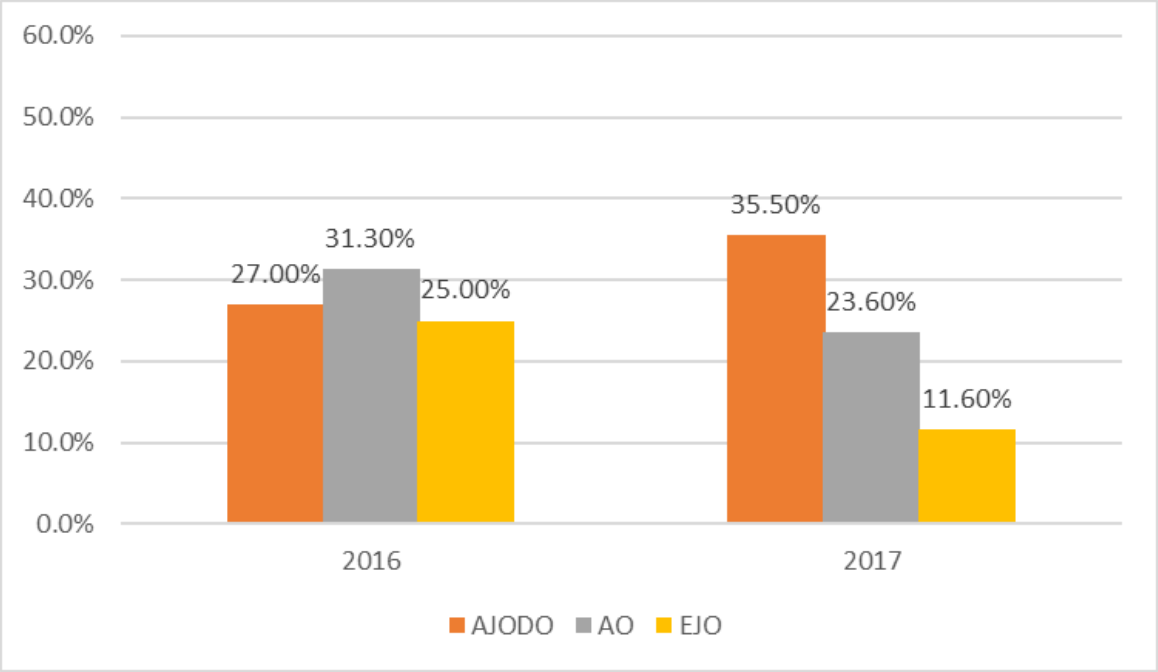
Of the eligible 599 articles meeting the inclusion and exclusion criteria, a total number of 162 published articles were deemed to have clustering effects and were eventually included in the final analysis. Of which, 88 of the included articles were published in 2016, while 74 articles were published in 2017.

The table below illustrates the numbers of articles included in the final analyses and percentage of articles included when compared to the total numbers of articles published in the respective journals for the year 2016 and 2017. Of these 162 articles from 2016-2017, a total of 79 were from the AJODO, 57 from the AO and 26 from the EJO. 31.2% of the total numbers of articles eligible articles from AJODO, 27.8% of the total number of articles AO, 18.4% of the total number from EJO were included in the final analyses. Overall, 27% of total number of eligible articles presented with clustering effects and therefore were included in the final analyses. The details on the type of articles published in respective journals have been discussed in later sections.
The details of the number of articles is displayed in detail in the table and graph below.

Table 5.3: Overall number and percentage of articles included in the final analysis based on journal and year of publication

	2016		2017		Total	
Journal	Total number of articles	Number of articles with clustering	Total number of articles	Number of articles with clustering	Total number of articles	Number of articles with clustering
AJODO	129	35 (27.0%)	124	44 (35.5%)	253	79 (31.2%)
AO	112	35 (31.3%)	93	22 (23.6%)	205	57 (27.8%)
EJO	72	18 (25.0%)	69	8 (11.6%)	141	26 (18.4%)
AJODO+AO+EJO	313	88 (28.1%)	286	74 (25.9%)	599	162(27.0%)

Figure 5.2: Percentage of articles included in the final analysis based on journal and year of publication.



5.2 Results of articles accounting for the clustering effects in the statistical analysis

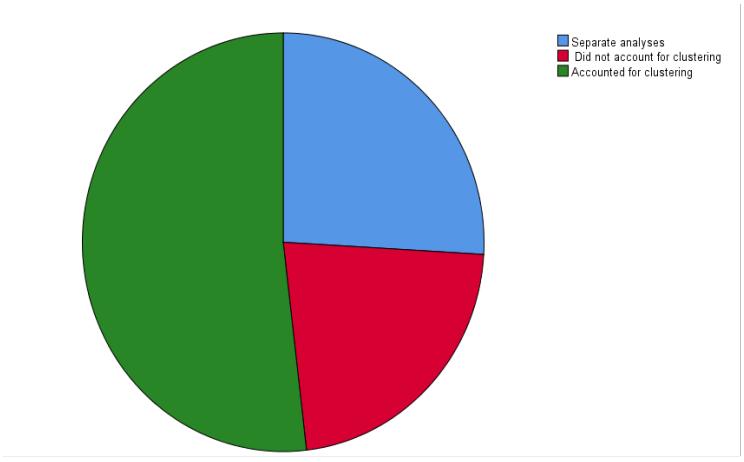
The statistical analysis of the 162 articles deemed to present with clustering effects in the study design were assessed thoroughly to determine if these effects were correctly accounted for in the statistical analysis. Of the included 162 articles with clustering effects, 84 (51.9%) of them correctly accounted for the clustering effects in the statistical analysis. The remaining 48.1% of the articles were subcategorised into articles which ignored the clustering effects and articles which analysed each outcome separately. 36 (22.2%) articles ignored the clustering effects in the statistical analysis, where the observations within a cluster were treated as if they were independent.

On the other hand, 42(25.9%) articles were categorised in the separate analyses group, where each observation or outcome within a cluster was treated as a separate variable.
For example, articles with measurements at multiple time points, with each time point analysed separately and articles with observations of multiple teeth on a periapical radiograph but each teeth was analysed separately. These articles were kept separate as the impact of not accounting for clustering is different in these two groups. This is further illustrated in the table and pie chart below.

Table 5.4: Frequency and percentage of articles accounted for clustering effects, ignored clustering effects and articles with separate analyses of each outcome

Accounted for clustering	Number of articles	Percentage (%)
Yes	84	51.9
No	36	22.2
Separate analyses	42	25.9

Figure 5.3: Distribution of studies which accounted for clustering, did not account for clustering effects and separate analyses.



5.3 Characteristics of the included articles and factors influencing accounting of clustering in statistical analysis

The articles included in this review can be further characterised based on journal of publication, type of study, region of authorship, collaboration with statistician, single or multicentre study, number of researchers, sample size, type of cluster and statistical significance. This is outlined in table 5.14. Also, further analysis has been done to determine the association of clustering with each characteristics.

5.3.1 Journal of Publication

The included articles were categorised according to the journals they were published in. 79 (48.8%), 57 (35.2%) and 26 (16%) articles published in the AJO-DO, AO and EJO journals respectively were considered to have clustering effects in the study design. This makes up a total of 162 articles with clustering effects in the study design.

The highest percentage of correctly accounting the clustering effects in the statistical analysis was found in the AJO-DO (57%), followed by AO (49.1%) and EJO (42.3%).

Table 5.5 illustrates the number of articles considered or did not consider the clustering effects in the statistical analysis, along with the number of articles flagged according to the journal of publication. Of the 79 articles published in the AJO-DO, 45 (57%) articles did account for the clustering effects in the statistical analysis, 22 (27.8%) ignored clustering and 12 (15.2%) articles did separate analyses. Almost half (n=28, 49.1%) of the included articles published in the AO accounted for the clustering effects in the statistical analysis. However, 7(12.3%) articles did not account for the clustering effects in statistical analysis and 22 carried out the separate analyses.

Of the 26 included articles from the EJO, 11 (42.3%) articles did account for the clustering effects in the statistical analysis. Half of the remaining articles did not account for the clustering effects and the other half were in the separate analyses category.

Table 5.5: Number of studies based on the journal of publication and its association with accounting for clustering

Journal Type	Accounted For Clustering			Total
	Separate analyses (n=42)(%)	Clustering ignored (n=36)(%)	Yes (n=84)(%)	
AJO-DO	12 (15.2%)	22 (27.8%)	45 (57%)	79 (100%)
AO	22 (38.6%)	7 (12.3%)	28 (49.1%)	57 (100%)
EJO	8 (30.8%)	7 (26.9%)	11 (42.3)	26 (100%)
Total	42 (25.9%)	36 (22.2%)	84 (51.9%)	162 (100%)

5.3.2 Type of Study

Each article was characterised to either observational or interventional study. Of the 162 included articles, observational studies made up 73.5% (n=119) of the total included articles over the two year period. As illustrated in the table below, the remaining 26.5% of (n=43) articles were interventional studies.

Table 5.6 illustrates the distribution of articles based on study type and its association with accounting for clustering effects. Little difference is observed in the number of articles accounting for the clustering effects between the interventional (n=21, 48.8%) and observational (n=63, 52.9%) studies. However, 32 (26.9%) of the observational studies and only 4 (9.3%) interventional studies did not take the clustering effects into consideration. Also, 24 observational studies and 18 interventional studies were categorised in the separate analyses group.

Table 5.6: Number of studies based on the type of study and its association with accounting for clustering

Study Type	Accounted for Clustering			Total
	Separate analyses (n=42)(%)	Clustering ignored (n=36)(%)	Yes (n=84)(%)	
Interventional	18 (41.9%)	4 (9.3%)	21 (48.8%)	43 (26.5%)
Observational	24 (20.2%)	32 (26.9%)	63 (52.9%)	119 (73.5%)

5.3.3 Region of Authorship

There was variation in the findings dependant on where the article originated. The region of authorship was recorded according to the country of origin of the first author.

For the number of articles based on region of authorship, Europe had the highest number of articles with 65 articles published with study designs that included clustering effects, followed by 51 articles from America and 40 articles from Asia. Articles that did not fall in any of the first three continents were grouped in the category, 'Other'. This included 6 articles originating from either Australia or New Zealand. The number and percentage of included articles published in each region can be seen in the table below. For accounting of clustering, articles in the category, 'Other' had the highest percentage, 66.7% (n=4) , followed by Europe, 53.8% (n=35), America, 51.0% (n=26) and the lowest, 47.5% (n=19) from Asia.

Table 5.7: Number of studies based on the region of authorship and its association with accounting for clustering

Region of authorship	Accounted For Clustering			Total
	Separate analyses (n=42) (%)	Clustering ignored (n= 36) (%)	Yes (n=84) (%)	
America	14 (27.5%)	11 (21.6%)	26 (51.0%)	51 (31.4%)
Asia	8 (20.0%)	13 (32.5%)	19 (47.5%)	40 (24.7%)
Europe	18 (27.7%)	12 (18.5%)	35 (53.8%)	65(40.1%)
Other	2 (33.3%)	0 (0 %)	4 (66.7%)	6 (3.7%)

5.3.4 Collaboration with statistician

The involvement of a statistician in analysis of the results of the included articles was determined by reviewing the authors' affiliation information in published articles. In the event the affiliation information in regard to the use of statistician was unclear, the author's name and university information was searched using Google to further clarify the involvement of a statistician.

The majority of the articles with clustering effects did not involve a statistician, making up 82.7% (n=134) of the included articles. However, we have only looked at the author lists in the articles and it is possible that a statistician could have been consulted but not listed as an author. Almost half (47.8%) of them did account for the clustering effect in the statistical analysis.

There were only 28 articles (17.3%) with statistician involvement in analysis of results.
The percentage of articles accounting for the clustering effects when having a statistician on board was higher than the articles without the presence of a statistician, as illustrated in table 5.8. There were only 2 (7.1%) articles with statistician involvement which did not account for the clustering effects in the statistical analysis. However, 6 out of the 28 (21.4%) articles did a separate analysis of analysing each variable separately.

Table 5.8: Number of studies based on the involvement of statistician and its association with accounting for clustering

Collaboration with statistician	Accounted for Clustering			Total
	Separate analyses (n=42) (%)	Clustering ignored (n= 36) (%)	Yes (n=84) (%)	
No	36 (26.9%)	34 (25.4%)	64 (47.8%)	134 (82.7%)
Yes	6 (21.4%)	2 (7.1%)	20 (71.4%)	28 (17.3%)

5.3.5 Multicentre study:

Overall, a higher proportion of articles with clustering effects were single centre studies, making up 92.6% of included articles. There were only 12(7.4%) multi-centre studies included in the final data analysis. This can be seen in the table below.

Clustering effects were correctly accounted for in 75% (n=9) of multicentre studies and 50.7% (n=76) of the single centre studies. It is worth noting none of the multicentre studies ignored clustering effects of the respective study.

24.0% (n=36) of the single centre studies did not take the clustering effects into considerations and 26.0% (n=39) were in the separate analyses group.

Table 5.9: Number of studies based on single or multicentre study and its association with accounting for clustering

Multicentre Study	Accounted for Clustering			Total
	Separate analyses (n=42) (%)	Clustering ignored (n= 36) (%)	Yes (n=84) (%)	
No	39 (26.0%)	36 (24.0%)	75 (50.0%)	150 (92.6%)
Yes	3 (25.0%)	0 (0.0%)	9 (75.0%)	12 (7.4%)

5.3.6 Number of authors reported

The number of researchers were grouped into the following categories:

1. One to three researchers
2. Four researchers
3. 5 and more than 5 researchers

Overall, more than half of the included articles (53.7%, n=87) involved five or more than 5 researchers, followed by four researchers (25.9%, n=42) and lastly, only 20.4% (n=33) of the included articles involved three or less researchers.

Of the 84 articles that have accounted for the clustering effects, 39.4% (n=13) of the articles involved three or less than three authors, followed by 52.4% (n=22) of the articles with four authors and 56.3% (n=49) of the articles with five or more than five authors.
This suggests that as the number of authors increased, the articles that accounted for clustering effects increased, as illustrated in Table 5.10.

Of the 36 articles which did not account for the clustering effects, 21.2% (n=7) of the articles involved three or less researchers, 35.7% (n=15) of the articles had four authors and 16.1% (n=14) articles with five or more than five authors.

Of the 42 articles, which did separate analyses, 39.4% (13) of the articles involved three or less authors, 11.9% (5) articles had four authors, 27.6% (24) articles involved five or more than five authors.

Table 5.10: Number of studies based on the number of researchers and its association with accounting for clustering

Number of researchers	Accounted for clustering			Total
	Separate analyses (n=42) (%)	Clustering ignored (n= 36) (%)	Yes (n=84) (%)	
<3	13 (39.4%)	7 (21.2%)	13 (39.4%)	33 (20.4%)
4	5 (11.9%)	15 (35.7%)	22 (52.4%)	42 (25.9%)
>5	24 (27.6%)	14 (16.1%)	49 (56.3%)	87 (53.7%)

5.3.7 Reporting of sample size

Articles are categorised as sample size reported when the sample size calculation is presented as the primary outcome and has statistical power to detect results that have a clinically meaningful difference.

The sample size calculation should ideally include the following components:[68]

i. The alpha: The value of alpha is most commonly 0.05. This means that there is a 5% chance of making a type 1 error, which is a false positive error.
ii. Power of the study: It is commonly 0.8, following 1 – β of 0.2. This means there is 20% chance of a type II error (false negative). It can also be interpreted as 80% probability of avoiding a type 2 error.
iii. The smallest effect of interest. It is defined as the minimal difference between the study groups that the investigator wishes to detect.
iv. The variance: The variability of the outcome measured is expressed as the SD in case of a continuous outcome. As the variance is an unknown quantity, investigators often use an estimate obtained from a pilot or previous study.

51.2% (n=83) of the included articles reported on the sample size calculation and 48.4% (n=79) did not. Of the 79 articles without sample size calculation, 39 (49.4%) articles accounted for the clustering effects in the statistical analysis, 21 did not and 18 articles did separate analyses of each outcome. Conversely, of the 83 articles, which did report on the sample size calculation, 45 (54.2%) articles accounted for the clustering effects in the statistical analysis, 15 (18.1%) did not and 23 (27.7%) articles did separate analyses.

A separate count of articles which accounted for the clustering effects in the sample size calculation was kept. Of the 83 articles, only 8 (9.6%) articles reported on the value of ICC or DE and accounted for the clustering effects in the sample size calculation.

Table 5.11: Number of studies based on the reporting of sample size and its association with accounting for clustering

Sample size reported	Accounted for Clustering			Total
	Separate analyses (n=42) (%)	Clustering ignored (n= 36) (%)	Yes (n=84) (%)	
No	19 (24.1%)	21 (26.6%)	39 (49.4%)	79 (48.8%)
Yes	23 (27.7%)	15 (18.1%)	45 (54.2%)	83 (51.2%)

5.3.8 Type of Cluster

This characteristic illustrates the rationale of including the articles in the final analysis. Table 5.12 displays the number of articles according to the type of cluster adopted in the study design. Silhouette studies where multiple participants rated the same image, were grouped in the 'multiple assessors' category. Studies involving multiple observations of teeth nested in the same individual were grouped in the 'multiple teeth' category. The 'multiple time point' category is made up of studies with multiple measurements from each subject at multiple pre-determined time points. Lastly, the category 'others' included studies such as TMJ assessments, TAD assessments and studies with geographical/ institution clusters.

Of the included 162 articles with clustering effects, 14 (8.6%) articles had clustering of multiple assessors and 46(27.7%) had clustering of multiple teeth. Most of them had clustering of multiple time points, compromising 92 (57.4%) of the 162 articles.

In regard to accounting for clustering, the group 'Others' had the highest percentage of 70%, followed by 64.3% and 51.1% respectively by the multiple assessors and multiple time point group. The multiple teeth group accounted for the least clustering with 45.7% only. It is essential to note that the most number (n=39, 41.9%) of articles with a separate analysis of the outcome were found in the multiple time point group.

Table 5.12: Number of studies based on the type of cluster and its association with accounting for clustering

Type of Cluster	Accounted for Clustering			Total
	Separate analyses (n=42) (%)	Clustering ignored (n= 36) (%)	Yes (n=84) (%)	
Multiple assessors	0 (0%)	5(35.7%)	9 (64.3%)	14 (8.6%)
Multiple teeth	3 (6.5%)	22 (47.8%)	21(45.7%)	46 (27.7%)
Multiple time points	39 (42.4%)	6 (6.5%)	47 (51.1%)	92 (56.8%)
Others	0 (0.0%)	3 (30.0%)	7 (70.0%)	10 (6.2%)

5.3.9 Statistical significance

Articles were considered to be statistically significant if the reported results of the primary outcome were found to be significant. In most articles, the outcome is thought to be significant when the P value is <0.05, unless stated otherwise. The significance of the results was compared to the level of significance set for the particular study. Also, if the results were considered statistically significant, it is usually reiterated in the conclusion of the articles. 109(67.3%) of the included articles concluded with statistically significant results. On the other hand, 53 (32.7%) articles reported the results were not statistically significant.

Majority of the articles (n=64, 58.7%) with statistically significant results did account for the clustering effects in the statistical analysis. However, 22 (20.2%) of the articles with significant results did not account for the clustering effects in the statistical analysis. This is worrying, as these articles might have incorrect conclusions.

Of the 53 articles with non-significant results, 20 (37.7%) did account for the clustering effects, 14 (26.4%) did not account for the clustering effects and 19 (35.8%) articles analysed each outcome separately.

Table 5.13: Number of studies based on the reporting of statistical significance and its association with accounting for clustering

Statistical significance	Accounted for Clustering			Total
	Separate analyses (n=42) (%)	Ignored clustering (n= 36) (%)	Yes (n=84) (%)	
No	19 (35.8%)	14 (26.4%)	20 (37.7%)	53 (32.7%)
Yes	23 (21.1%)	22 (20.2%)	64 (58.7%)	109 (67.3%)

Table 5.14: Distribution of the 162 articles with clustering effects based on journal of publication, type of study, region of authorship, collaboration with statistician, single or multicentre study, number of researchers, sample size, type of cluster and statistical significance.

Variables	Category	Total, N (%)	Clustering ignored, N (%)	Accounted for clustering effects N (%)	Separate analyses, N (%)
Journal of Publication	AJO-DO	79 (100.0)	22 (27.8)	45 (57.0)	12 (15.2)
	AO	57 (100.0)	7 (12.3)	28 (29.1)	22(38.6)
	EJO	26 (100.0)	7 (26.9)	11 (42.3)	8 (30.8)
Type of Study	Interventional	43 (100.0)	4 (9.3)	21 (48.8)	18 (41.9)
	Observational	119 (100.0)	32 (26.9)	63 (52.9)	24(20.2)
Region of authorship	America	51 (100.0)	11 (21.6)	26 (51.0)	14 (27.5)
	Asia	40 (100.0)	13 (32.5)	19 (47.5)	8 (20.0)
	Europe	65 (100.0)	12 (18.5)	35 (53.8)	18 (27.7)
	Other	6 (100.0)	0 (0%)	4 (66.7)	2 (33.3)
Collaboration with statistician	No	134 (100.0)	34 (25.4)	64 (47.8)	36 (26.9)
	Yes	28 (100.0)	2 (7.1)	20 (71.4)	6 (21.4)
Multicentre study	No	150(100.0)	36 (24.0)	75 (50.0)	39(26.0)
	Yes	12 (100.0)	0 (0.0)	9 (75.0)	3 (25.0)
Number of researchers	<3	33 (100.0)	7 (21.2)	13 (39.4)	13 (39.4)
	4	42 (100.0)	15 (35.7)	22 (52.4)	5 (11.9)
	>5	87 (100.0)	14 (16.1)	49 (56.3)	24 (27.6)
Reporting of sample size	No	79 (100.0)	21 (26.6)	39 (49.4)	19 (24.1)
	Yes	83 (100.0)	15 (18.1)	45 (54.2)	23 (27.7)
Type of Cluster	Multiple assessors	14 (100.0)	5 (35.7)	9 (64.3)	0 (0.0)
	Multiple teeth	46 (100.0)	22 (47.8)	21 (45.7)	3 (6.5)
	Multiple time points	92 (100.0)	6 (6.5)	47 (51.1)	39 (42.4)
	Others	10 (100.0)	3 (30.0)	7 (70.0)	0 (0.0)
Statistical significance	No	53 (100.0)	14 (26.4)	20 (37.7)	19 (35.8)
	Yes	109 (100.0)	22 (20.2)	64 (58.7)	23 (21.1)

5.4 Summary of statistical methods used in articles presenting with clustering effects in the study design

The table below displays the frequency and the percentage of the statistical methods used in the 162 articles included in this review. It also includes information on the frequency of the statistical method used in all articles flagged for clustering including those which accounted and did not account for the clustering effect.

The most commonly used statistical method was mixed models, which was noted in 43.8% (n=71) of the included articles. This was followed by the t-test (33.3%, n=54), ANOVA (11.7%, n=19) and Chi square (4.9%, n=8) methods. Only 1.9% (n=3) of the included articles performed logistic regression as a statistical method. Linear regression and survival category was performed in 1.2 % (n=2) of the included articles respectively. However, none of the articles used correlations as a statistical method. Lastly, 1.9% (n=3) of the included articles did not report on any statistical methods used.

Of the included 162 articles, only 84 (51.9%) articles correctly accounted for the clustering effects in the statistical analysis. All the articles, which used mixed model as a statistical method, did correctly adjust for the clustering effects in the statistical analysis. This is followed by Chi square (25.0%), ANNOVA (21.1%) and the lowest was in the t-test category (13.0%).

Of the 36 articles that did not address the clustering effects, 14 used the t-test, 11 used ANOVA method, 6 used Chi-square and only 1 article reported the use of logistic regression. There were two articles using survival category and they both did not take the clustering effects into consideration during statistical analysis. Two articles did not even have a statistics category.

In total, there are 41 articles in the separate analyses category. A large number of these articles (n=33) were from the t-test group and four articles used ANOVA as a statistical method. Two articles used linear regression and logistic regression respectively. Only one article used no statistical category.

Table 5.15: Frequencies and percentages of statistical methods used in articles which accounted, did not account for clustering effects in statistical analysis and articles with separate analyses.

Statistical group	Accounted for clustering			Total
	Separate analyses	No	Yes	
Mixed models	0 (0%)	0 (0%)	71 (100%)	71 (43.8%)
t-test category	33 (61.1%)	14 (25.9%)	7 (13.0%)	54 (33.3%)
ANNOVA	4 (21.1%)	11 (57.9%)	4 (21.1%)	19 (11.7%)
Chi-square	0 (0%)	6 (75.0%)	2 (25.0%)	8 (4.9%)
Logistic regression	2 (66.7.%)	1(33.3%)	0 (0%)	3 (1.9%)
No statistics category	1 (33.3%)	2 (66.7%)	0 (0%)	3 (1.9%)
Survival category	0 (0%)	2 (100%)	0 (0%)	2 (1.2%)
Linear regression	2 (100%)	0 (0%)	0 (0%)	2 (1.2%)
Correlations	0 (0%)	0 (0%)	0(0%)	0
Total	41 (25.3%)	36 (22.2%)	85 (52.5%)	162 (100%)

5.5 Univariable and multivariable logistic regression for articles accounting versus non-accounting for clustering effects when including 'separate analyses' category articles

The results of univariable and multivariable model produced when including journal of publication, region of authorship, collaboration with statistician, single or multicentre study, type of study, number of researchers and sample size reporting as predictors are outlined in Table 5.16. Articles belonging to the separate analyses category and articles non-accounting for clustering effects category were combined. The variable which had a statistically significant effect on accounting for the clustering effects in the statistical analysis was the involvement of statistician. (unadjusted odds ratio= 2.73; p= 0.026; 95% CI: 1.13-6.64). The interpretation of the univariable logistic regression (Table 5.16), show that the odds of correctly accounting for the clustering effects of the study design in the statistical analysis was 2.73 times greater with the involvement of a statistician in the study. The variable single or multicentre study was also included in the multivariable model as the cut of point to be included in a multivariable model is 0.1. The results of multivariable analysis shows the collaboration of statistician as a significant predictor for accounting of clustering, where the odds of accounting of clustering was 2.91 times greater when having a statistician involved in a study.

Table 5.16: Univariable and multivariable logistic regression-derived odds ratios (ORs) and confidence intervals (CIs) for articles accounting versus non-accounting for clustering effects, when including the separate analyses category articles in the not accounted for clustering effects category. [accounted vs non accounted for clustering effects (ignored clustering + separate analyses)

Variable	Category	Univariable analysis			Multivariable analysis		
		OR	95% CI	p-value	OR	95% CI	p-value
Journal	AJODO		Baseline				
	AO	0.73	0.37, 1.45	0.366			
	EJO	0.55	0.23, 1.36	0.197			
Region of authorship	America		Baseline				
	Asia	0.87	0.38,1.99	0.742			
	Europe	1.00	0.47,2.12	0.991			
	Other	2.40	0.67,8.67	0.180			
Collabration with statistician	No		Baseline				
	Yes	2.73	1.13,6.64	0.026	2.91	1.19, 7.11	0.019
Single/Multicentre study	No		Baseline				
	Yes	3.00	0.78, 11.52	0.109	3.37	0.87, 13.09	0.079
Study Type	Interventional		Baseline				
	Observational	1.18	0.59, 2.37	0.645			
Number of researchers	<3		Baseline				
	4	1.69	0.67, 4.27	0.265			
	>5	1.98	0.88, 4.49	0.100			
Sample size calculation reported	No		Baseline				
	Yes	1.22	0.66, 2.25	0.537			

5.6 Univariable and multivariable logistic regression for articles accounting versus non-accounting for clustering effects in statistical analysis

A direct comparison of the articles accounting for the clustering effects and non-accounting for the clustering effects was done. In this analysis, articles from the separate analyses category were excluded.

Table 5.17 depicts the results of the univariable and multivariable produced, which includes journal of publication, region of authorship, collaboration with statistician, single/multicentre study, type of study, number of researchers and sample size calculation as predictors. The only variable which had a statistically significant effect on the accounting of clustering in the statistical analysis is the involvement of statistician (unadjusted odds ratio= 5.31; p= 0.030; 95% CI: 1.17-24.09). Journal and region of authorship were also included in the multivariable model after backward elimination, although were not significant at the 5% significance level. The adjusted odds ratio of accounting for clustering when having a statistician on board is 8.20 (p=0.010, 95%CI: 1.65- 40.83).
The significance of the variable single or multicentre study as a predictor could not be estimated, as there was no multicentre study that did not account for the clustering effects.

Table 5.17: Univariable and multivariable logistic regression-derived odds ratios (ORs) and confidence intervals (CIs) for articles accounting versus non-accounting for clustering effects in the statistical analysis.

Variable	Category	Univariable analysis			Multivariable analysis		
		OR	95% CI	p-value	OR	95% CI	p-value
Journal	AJODO		Baseline				
	AO	1.96	0.74, 5.17	0.177	1.76	0.63, 4.91	0.280
	EJO	0.77	0.26, 2.25	0.631	0.28	0.07, 1.13	0.073
Region of authorship	America		Baseline				
	Asia	0.62	0.23, 1.68	0.345	0.63	0.22, 1.83	0.395
	Europe	1.12	0.42, 3.00	0.829	1.61	0.52, 5.02	0.411
	Other	4.23	0.48,37.17	0.193	7.31	0.72, 73.91	0.092
Collabration with statistician	No		Baseline			Baseline	
	Yes	5.31	1.17,24.09	0.030	8.20	1.65, 40.83	0.010
Single/Multicentre study							
Study Type	Interventional		Baseline				
	Observational	0.38	0.12, 1.19	0.085			
Number of researchers	<3		Baseline				
	4	0.79	0.26, 2.44	0.682			
	>5	1.89	0.63, 5.63	0.256			
Sample size calculation reported	No		Baseline				
	Yes	1.62	0.73, 3.56	0.234			

5.7 Effects of clustering on finding significant results

In the logistic regression above, statistical significance was not included as part of determining the predictors of correct handling of clustering effects. This is because, we believe the statistical significance of the results are derived after conducting the study.

The interpretation of the univariable logistic regression show that, when including the articles with separate analyses, the odds of correctly accounting for the clustering effects of the study design in the statistical analysis was 2.35 times greater in articles with significant results. (unadjusted odds ratio= 2.35; p= 0.013; 95% CI:1.20-4.60).

When excluding the articles with separate analyses, we found the odds of accounting for clustering effects was 2.04 times greater in articles presenting with significant results.

(unadjusted odds ratio= 2.04; p= 0.096; 95% CI: 0.88-4.70).

Table 5.18: Univariable logistic regression-derived odds ratios (ORs) and confidence intervals (CIs) on statistical significance and accounting clustering effects when including and excluding the separate analyses category articles

Category	Statistical significance	OR	95% CI	p-value
Adjusted for clustering when including the separate analyses articles	Significant			
	No		Baseline	
	Yes	2.35	1.20, 4.60	0.013
Adjusted for clustering when excluding articles with separate analyses	Significant			
	No		Baseline	
	Yes	2.04	0.88, 4.70	0.096

5.8: Inter and Intra reliability assessment:

Four issues of the journals were reassessed after three months into data collection. This included two issues from AJODO and one each from EJO and AO. The kappa score for intra-examiner reliability was 0.913 indicant an excellent reliability during data extraction. Compared to the initial data collection, two additional articles were identified as presenting with clustering effects and disagreement on five boxes on the variables were noted. This shows the possibilities of including extra articles was higher than missing on articles presenting with clustering effects. After initial shortlisting by BB, all the articles were discussed with my supervisor (GB) before including in the final analysis. Therefore, it was decided that no formal analysis to be carried out to assess inter-rater reliability.

Chapter 6: Discussion

6.1 Summary of the main findings:

A total number of 162 published articles with clustering effects in the study design met our eligibility criteria and were included in this study. This resulted from a search of all the articles published in the AJODO, EJO and AO journals, in 2016 and 2017.

When exploring the articles in detail, clustered study designs were encountered in articles under the following circumstances. Where multiple observations of several sites were collected from each subject, repeated measurements at pre-determined time points, when multiple participants rated the same image and in articles presenting with institutional clusters.

This study found only 84 (51.9%) of the included articles, correctly accounted for the clustering effects in the statistical analysis. This suggests a potentially poor awareness of clustering effects among researchers, as approximately half of the articles did not take the clustering effects into considerations.

36 (22.2%) articles ignored the clustering effects in the statistical analysis, where the observations within a cluster were treated as if they were independent and analysed as a single outcome. Failure to account for the clustering effects in the statistical analysis can result in increase of sample size, artificial reduction of standard error, leading to p-values which are too small.

42(25.9%) articles were categorised in the separate analyses group, where each observation within a cluster were analysed as separate outcomes. As illustrated in Table 5.12, majority of the articles (n=39, 93%) in the separate analyses were from the multiple time point group.
This included articles with repeated measurements collected at pre-determined time points and observations at each time points were analysed separately. This multiple significance test on a data set increases the probability of a Type I error, finding a statistically significant result even if the null hypothesis is true, just by chance alone.

For example, Alsayed Hasan et al (2017)l conducted a study aiming to evaluate the effectiveness of low-level laser therapy in accelerating orthodontic tooth movement. This was a two-arm randomised controlled trial, where patients were allocated to either the laser or the control group. In both groups, patients had extraction of the upper first pre-molars and the tooth movement of the crowded maxillary incisors were assessed. Patients in the laser groups received the laser treatment at pre-determined time points until the end of the aligning and levelling treatment phase. Alignment

progress was evaluated on the study casts which was taken at four time points including, before inserting the first archwire (T0), after 1 month of treatment commencement (T1), after 2 months (T2), and at the end of the leveling and alignment stage (T3). The outcome measures were the overall time needed for leveling and alignment and the leveling and alignment improvement percentage. A two-sample t-test was applied to evaluate the differences of the outcomes in each studied time point between the two group.[68] Because the improvement in the levelling and alignment was analysed at each time point as a separate outcome rather than analysed as repeated measures, the measurements from the subject could not be regarded as a cluster. However, this potentially increases the chance of a Type 1 error due to multiple hypothesis testing, unless a statistical correction was applied.

It is particularly concerning when results are interpreted solely based on p-values to derive conclusions. Pandis pointed out that focusing on p-values might be misleading as it does not provide sufficient information about the effect size of a treatment. Rather a p-value, on its own, only provides the strength of the evidence against the null hypothesis.[52] Nevertheless, ρ-values are influenced by sample size and standard deviation. Thus, a small ρ-value does not necessarily indicate a large intervention effect and vice versa.

Instead, researchers should place emphasis on the effect estimate of the study, as they provide more information on the treatment effect. This parameter be in terms of confidence intervals, difference in mean, odds ratio, proportion, etc. If the 95% CI of the effect estimate contains the value 0, this means that the p-value will be greater than 0.05.[56] Conversely, if the 95% CI does not contain the value 0, then the p-value will be strictly less than 0.05.[56] Odd ratio represents the odds of the occurrence of the outcome of interest given a particular exposure. When using odds ratio, the situation of no difference will be indicated by the value of 1 instead of 0. An odds ratio less than 1 suggests that the effects of treatment are less likely to occur, given a particular exposure. Whereas, an odds ratio greater than 1 suggests an association between both events, and the treatment effects are more likely given a particular exposure. Hence, if the 95% CI of the ratio contains the value 1, the p-value will be greater than 0.05.[56] Alternatively, if the 95% CI does not contain the value 1, the p-value is strictly less than 0.05.[56] This shifts the interpretation of results from either a significant or non-significant approach to the size and range of the effect which offers valuable information when evaluating evidence to make a clinical decisions.[50,57, 58]

The statistical methods used to account for the clustering effects are displayed in table 5.15. As discussed above, the two main approaches to the analysis of clustered trials are cluster level analysis

and individual level analysis. A relatively high percentage (43.8%) of articles with mixed model was highlighted in this study. Off the 84 articles which accounted for the clustering effects in the statistical analysis, 71 (83.5%) articles used the mixed models method and 4 (4.7%) articles were from the ANOVA category. These models of analysis were appropriate for evaluating the correlated data as they allowed the evaluation of individual results while simultaneously accounting for the clustering effects. The remaining nine articles conducted the cluster level analysis. Seven articles were from t-test and two articles from the chi-square category. Here, statistical analysis was conducted at the patient level. A summary of the outcome was measured for each cluster followed by statistical analysis comparing the effects estimate between the treatment arms.

Factors influencing whether an article correctly accounted for the clustering effects were also examined. The included articles were investigated with regards to various trial characteristics which included the following variables:

- Journal of publication
- Region of authorship
- Collaboration with statistician
- Single/Multicentre study
- Study type
- Number of researchers

The results of the Univariable and multivariable analysis are depicted in Table 5.16 and 5.17.

i. When including the articles with separate analyses in the logistic regression, a significant association of the collaboration of statistician (unadjusted odds ratio= 2.73; p=0.026; 95% CI: 1.13-6.64) with correctly accounting of the clustering effects in the statistical analysis was found. This reflects having a statistician as one of the authors may help in appropriate management of the statistical aspects in a study. A statistician on board could provide some professional advice along with accurate statistical reporting. The multivariable model reveal the same variables of collaboration with statistician (adjusted odds ratio: 2.91; p=0.019; 95% CI: 1.19-7.11) as significant predictor in accounting for the clustering effects.

ii. When excluding the articles with separate analyses, from the logistic regression and having a comparison of articles accounting versus non-accounting for clustering effects, the only variable with significant association of correctly accounting for the clustering effects was the involvement of statistician (unadjusted odds ratio=5.31; p=0.030; 95% CI:1.17-24.09). The

adjusted odds ratio of accounting for clustering when having a statistician on board is 8.20 (p=0.010, 95% CI 1.65-40.83). However, it has a wide confidence interval suggesting less precise results. This is similar to the above finding, where involvement of a statistician was found to have a significant correlation with accounting for clustering.

6.2 Summary of characteristics of the included articles

The articles included in this study were from the three major orthodontic journals, including AJODO, AO and EJO. Handsearching of the content of the issues published in 2016 and 2017 was done. These are the main journals widely read in Europe. Furthermore, the selection of the four orthodontic journals namely AJODO, AO, EJO and JO were also recommended by Shimada et al for practice of evidence-based orthodontics in order to gather high quality material related to orthodontics.[69] The rational for selecting only this three journals (AJODO, AO, EJO) was to have the similar sample with the study done by Koletsi et al (2012). This allowed to draw a comparison with the previous study done by Koleksi et al(2012), and further assess if there is a change of orthodontic studies which account for clustering in statistical analysis. However, due to the time constraint in this study, only issues published in the 2016 and 2017 was included in this study, making a total of 48 journals. On the other hand, the study by Koletsi et al(2012) included the most recent 24 issues of each journal from December 2010 backwards. Thus, a total of 72 issues were included in the study by Koletsi et al(2012), having a larger sample compared to our study.

As depicted in table 5.1, half of the articles identified during the initial screening were from AJODO, mainly because it is a journal published monthly as compared to AO and EJO which is published bimonthly. The overall number of articles included in the final analysis according to the journals they were published in are displayed in Table 5.5. 79 (48.8%), 57 (35.2%) and 26 (16%) articles published in the AJODO, AO and EJO journals respectively were considered to have clustering effects in the study design.

When examining the study type, 73.5% of the included article were observational studies. It is worth to note, majority (92.6%) of the included articles were single centre studies. This is similar to the distribution of the articles in Koletsi et al study(2012) , which reported of 63.5% of observational studies and 80% were single centre studies included in the final analysis.

17.3% of the articles had a statistician involved in the study. The involvement of a statistician was determined by reviewing the authors' affiliation information and acknowledgements in published

articles. As illustrated in table 5.8, the percentage of articles accounting for the clustering effects when having a statistician on board was higher than articles without the presence of a statistician. Additionally, this was the only variable found to have a significant association with correctly accounting for clustering effect in the logistic regression analysis when including as well as excluding the articles with separate analyses. Papageorgiou(2019) reported statisticians are more likely to be involved in orthodontic trials compared to periodontic trials.[70]

Considering the information collected on the number of researchers, more than half of the included articles (53.7%, n=87) involved five or more than 5 authors, followed by four authors (25.9%, n=42) and lastly, less than three authors only made up 20.4% (n=33) of the included articles.

Articles were also characterised according to the statistical significance. 67.2% of the articles reported of significant results and 32.7% with non-significant results. From the 109 articles which reported of significant results, 22(20.2%) of them did not consider the clustered study design in the analysis. This arises the question of the validity of the results. How many of these studies with significant results which did not account for the clustered design might have had non-significant results if the clustered designs was considered? As discussed above, the over-dependence on p-value and this incorrect handling of the clustering effects could potentially result in false positive results and incorrect conclusions. Furthermore, it is conceivable most orthodontic journals prefer on reporting of significant results. Koletsi et al(2012) found an association between impact factor and statistically significant results where journals with impact factor had a 100% increased probability of publishing articles with significant results compared with journals with no impact factor.[64] This has led most authors to emphasizes on significant findings in their result.

Ideally, the clustered study design should be taken into consideration during the sample size calculation. The reporting of sample size acts as an indicator as to whether the researcher has adequately designed the study in advance. Whether or not it is clustered, and have taken all factors into consideration. The CONSORT and STROBE guidelines have emphasised on the importance of accurate reporting of the method of sample size calculation in a cluster RCT and longitudinal study, respectively.[67, 71]

In 2013, Koletsi et al published a review that analysed the quality of reporting of sample size calculation in RCTs published in eight leading Orthodontic journals.[72] Off the 139 RCT's identified, only 41(29.5%) articles reported complete and feasible sample size calculations while the majority (70.6%)

of the included studies failed to report on the sample size calculation.[72] Similarly, Pandis et al reported of only 7% of the included articles from the six major dental specialty journals provided the sample size calculations.[73] In the medical literature, Elridge et al reported of a review of cluster randomised trials published from 1997 to 2000, where 20% of published trials accounted for clustering in sample size calculation and 59% of published trials accounted for clustering in analyses.[73]

In this present study, 51.2% (n=83) of the included articles reported on the sample size calculation in the methodology section. However, only 8 of the articles accounted for the clustered study design in the sample size calculations by appropriate reporting of the value of ICC or the DE. As discussed in the literature review, the correlated nature of the data should be taken into account during sample size calculation in a clustered trial. Failure to do so, results in an under power study and incorrect inferences. Thus, the 'design effect' (DE) can be used to estimate the extent to which the sample size should be inflated to accommodate for the similarity of this clustered data.

Data on the type of cluster was collected to justify the rationale of including these 162 articles in this study. Of the included 162 articles with clustering effects, 14 (8.6%) articles had clustering of multiple assessors, 46(27.7%) had clustering of multiple teeth and 10 (6.2%) articles were from the 'Other' category. Most of them had clustering of multiple time points, making up 92 (56.8%) of the 162 articles. Also, 39 of the 41 articles from the separate analyses group belong to the multiple time point category and 3 from the multiple teeth category.

6.3 Comparisons of findings with previous published research:

Four other reviews were found which examined articles on clustering effects. Two of which were from the dental literature (Koletsi et al 2012, Felming et al 2013)[1, 8] and two from the medical literature (Martin Bland 2004, Eldridge et al 2004)[5,73] . Therefore, the findings were compared to these previous similar studies.

The present study had a similar approach to the review by Koletsi et al(2012) and Fleming et al(2013) which hand searched the selected journals to identify papers with clustered designs. Koletsi et al hand searched the most recent 24 issues of the AJODO, AO and EJO from December 2010 backwards and concluded only a quarter of the included studies, where clustering was evident, accounted for the clustered designs in the statistical analysis. Mixed models and repeated ANOVA were the most commonly used statistical methods in the articles accounting for the clustering effects. Additionally, they found an association between type of journal and accounting of clustering, where articles published in AO were more likely to correctly account for the clustered designs in the analysis.[1] Fleming et al investigated clustered design articles in not only orthodontic journals but in the five leading dental specialty journals. This included journals in Orthodontics, Endodontology, Maxillofacial, Periodontology and Paediatric Dentistry.[8] They reported of 39.1% of the included studies with clustered designs that addressed the clustering effects appropriately.[8] The commonly used statistical methods in these articles were mixed models followed by t-test and lastly analysis of variance (ANOVA). This study found few factors influencing the accounting of clustering. This includes the similar factor reported by Koletsi et al(2012), type of journal. Additionally, the continent of origin and number of researchers were also found to be significant predictors. Better statistical management of clustering effects were found in Periodontology journals, articles published by European researchers and with greater numbers of authors.[8]

In contrast, this present investigation showed a slightly higher percentage of articles accounting for the clustering effects when compared to the above two discussed studies. We found 51.9 percent of the included articles accounted for the clustered design in the statistical analysis. Furthermore, there was a separate list of articles which analysed each of the outcome separately. Articles with data collected at repeated time points and accounted for each of the time points separately were kept in this 'separate analyses' group. We were unable to determine which category were these similar types of articles included in the Koletsi (2012)and Fleming(2013) study as there was no information available on the type of clustered articles included in the Koletsi(2012) and Fleming(2013) studies. Neither there was any effort taken to contact the authors. Also, the sample size in this present study

was smaller than the above mentioned studies. Although we selected the similar journals included in the study conducted by Koletsi et al (2012), we only looked into issues published over the two year period, 2016 and 2017. However, it should be noted that the interpretation made in this context may have biased the findings as there was inherent subjectivity in interpreting each article conclusions. The statistical methods used in the included articles accounting for the clustered designs were almost the same to the Koletsi and Fleming studies, with mixed models being the most commonly used statistical analysis.

In contrast to the study by Koletsi et al and Fleming et al, this investigation collected information of two additional variables including the reporting of sample size and type of cluster. Knowing the type of cluster in these articles provides justification of including the article in the study. According to CONSORT and STROBE guidelines, the reporting of sample size is a requisite for inclusion. It also allowed the examiner (BB) to take note of articles which considered the clustered study design in their sample size calculations. However, the only factor which was significantly associated with accounting for the clustering effects when including and excluding the articles of the 'separate analyses' group was the involvement of a statistician. This factor did not match with either one of the studies which found associations with the type of journal (Koletsi et al & Fleming et al), number of authors (Fleming et al) and continent of authorship (Fleming et al).[1, 8]

The common type of cluster employed is the medical literature is the institutional cluster. This is because most of the trials involved practices and this group of patients within the general practice setting forms a cluster. However, this present study did not find any articles of the institutional cluster type. Donner et al investigated on the methodological features and statistical analysis of non-therapeutic intervention trials employing cluster randomisation design.[30] Cluster trials published in the medical and epidemiological literature from January 1979 to august 1989 were included.[30] Sixteen articles were identified and only half of them used appropriated statistical methods to account for the clustering effects.[30] Eldridge et al conducted a systematic review of cluster randomised trials in the medical literature. Electronic and hand searching of cluster randomised trials involving primary health care published from 1997 to year 2000 was done. 152 published cluster randomised trials were included in the final analysis. 59 percent of them correctly accounted for the clustering effects in the statistical analysis.[73] This shows there has been a rise in the number of cluster trials involving primary health care over the years. Fortunately the quality of study designs and reporting has improved as well. The percentage of articles accounting for the clustered study design in the statistical analysis increased when compared to the study by Donner et al(2000).

6.4 Limitations of the study

6.4.1 Design of the study

The foundation of this study was mainly based on the previous published literature looking at clustering effects with further refinement made to meet the aim and objectives of this study.

6.4.2 Inclusion and exclusion criteria

This study had stringent inclusion and exclusion criterion to ensure appropriate articles with clustering effects were included in the final analysis. This study limited the inclusion of the studies to those published in the selected three journals which is AJODO, AO and EJO. This limited generalisability could have potentially underestimated the number of studies associated with clustering effects in the orthodontic literature and perhaps introduced selection bias. Also, these selected journals were not the top three ranked journals based on the 2017 SJR indicator. However, they were selected mainly to have the similar sample with the study done by Koletsi et al(2012) and enable the examiners to draw a comparison.

Furthermore, study involving animals, *in vitro* and laboratory studies were excluded from this study. This potentially could result is skewing of the number of articles published in the respective journals and affect the results. However, we decided so as the purpose of this study was to determine the clustering effects in clinical studies involving patients.

Finally, this study was limited to the selected English texts journals only. This may have introduced some language bias into the study. Some studies have shown that researchers are more likely to publish in non-English-language journals if the results are negative and in English language journals if they are positive. This phenomenon was demonstrated in the German literature by Egger et al, where the results showed only 35% of trials published in German had produce significant results compared to 63% of the articles published in the English literature.[15, 75] Including non-English-language articles within this study would have required collaboration between other parties to help in the translation of the articles. Hence, this was not felt appropriate within the remit of this project.

6.4.3 Identification of papers

A systematic review would commonly include articles from a wide range of databases. However in this study, only hand searching of the selected three journals was performed. The restriction of only selecting articles published in the specific journals over the two-year period only would have resulted in a lower percentage of clustering articles within those reviewed. When comparing to the other

similar studies in the dental literature, most of the studies looked into more issues of publication. The study by Koletsi et al(2012) reviewed 24 issues of each of the journal and the study by Fleming et al(2013) included 30 issues of each of the included journals. Therefore, this study sample is comparatively smaller, but by no means inferior.

This was a retrospective, observational study that fundamentally was open to bias. There is a possibility of mistakes which were made due to human errors in which articles which should have been included were unintentionally omitted. Therefore, precautions were taken by conducting two cycles of data collections for each of the journals.

Additionally, like most reviews there is an inevitable subjectivity in the screening process, especially when there are few researchers involved and many articles to be assessed. To prevent any inconsistence of the screening, a pilot study was done between the first author (BB) and research supervisor (GB) prior to the commencement of the data collection. This allowed to identify potential problems as well as improving consistency and precision of results.

6.4.4 Data extraction and analysis

In view of the large volume of articles screened and information on variables to be extracted from the included articles, there are possibilities of reporting bias by the first examiner (BB). Hence, not more than two issues were assessed at a time. Also, all of the articles were later discussed with the research supervisor (GB) during research meeting for confirmation and further assessment on the statistical analysis. However, there was no effort made to contact the authors of the included studies for study clarification especially on the statistical methods as this was beyond the remit of this study. Furthermore, the assessment of the clustering effects of the articles was hinged purely based on what was reported in the included articles. However, lack of reporting does not necessarily indicate that provisions regarding clustering effects were not made during study design and data analysis stage.

6.4.5 Quality

The standard quality assessment of the articles such as Cochrane bias tool was not done in this present study. The methodological quality of the study was determined only to the extent of determining if clustering was adopted in the study design followed by interpretation of the results.

6.4.6 Reliability

The kappa score for intra-rater reliability was 0.913, indicating excellent intra-rater reliability. However, the inter-rater reliability was not assessed as all the articles included in the study were discussed with the study supervisor (GB) who is a statistician. Having a statistician as a supervisor is the strength of this study. A second opinion was sought to ensure accurate data extraction and precise interpretation of the statistics in the articles presenting with clustered study designs.

6.5 Research implications

This review shed some light onto the challenges associated in the design, conduct and analysis of clustered studies in the orthodontic literature. Discounting of the clustering effects in the statistical analysis leads to incorrect study results, and this has important implications on study conclusions. As suggested in the previous published literature, the following key recommendations have been made for both authors and readers.

6.5.1 Author strategies:

The following suggested strategies are for the authors to consider when adopting a clustered study design:

i. Consider the methodological issues of a clustered study at the planning phase of the study. Researcher should determine and include the sample size calculations, ethical considerations, outcome of study and choice of analysis approach prior to the start of trial.
ii. The reliable reporting of trials can be improved by adhering to the CONSORT and STROBE guidelines for cluster RCTs and observational studies, to ensure all the relevant information is provided.
iii. Pre-specify a primary outcome at the study design stage
iv. Publish the estimate of intracluster correlation or between cluster variation to allow other researchers use this information in calculating sample size when planning further studies.
v. Perform formal and appropriate cluster or individual level analysis to correctly account for the clustering effects in the statistical analysis
vi. Use of confidence intervals when reporting results, instead of only reporting p-values with emphasis on either significant or non-significant results only.[75,73,53]

6.5.2 Readers strategies:

Readers should have a good understanding on the various types of study designs and be able to identify a clustered study design. Being able to critically appraise an article allows one to assess the methodology, analysis and interpretation of the results systematically. With a good knowledge on statistics, one should be able to determine if the clustering effects are correctly addressed in the statistical analysis. Significant results when not accounting the clustering effects should be interpreted cautiously. Furthermore, readers are recommended to give more importance on the confidence intervals instead of p-values.

6.6 Direction for future research

The following recommendations have been made for future research.

i. To repeat in few years to assess the extent these clustering effects are correctly accounted for in the statistical analysis.
ii. To explore the search in articles published in non-English language to reduce the risk of language bias
iii. To explore on the articles presenting with clustered design that did not account for the clustering effects and yet reported with significant results. It would be interesting to note how many of these articles with significant results would have become non-significant if the clustering effects were accounted correctly in the statistical analysis.
iv. Consider communicating with authors of the previous similar studies, to determine which types of articles were included in the study. This would be particularly helpful in assessing articles with multiple time points which were assessed as separate outcome variables or multiple measurements on the same subjects which are analysed separately too. In this study, we have made a separate flagged list of these articles. However, it be interesting to note how the authors of the previous studies categorised these articles.

Chapter 7: Conclusion

1. Articles presenting with clustering effects are commonly encountered in orthodontic journals.
2. A total of 162 articles published over the two-year period were included in this review with 51.9% of the articles (n=85) correctly accounted for the clustering effects in the statistical analysis.
3. The majority(84%) of the articles which accounted for clustering used mixed models, which are flexible methods allowing for modelling of clusters as random effects. The choice of statistical method used to account for the clustering effects would depend on the research design.
4. The only significant factor influencing accounting for the clustering effects when including and excluding the articles with separate analyses done, was the involvement of a statistician. It is advisable to involve a statistician in a cluster study to ensure the methodological and statistical issues are addressed appropriately.
5. Not accounting for the correlated data in a cluster trial can lead to incorrect inferences which may have an implication on our clinical practice.
6. In contrast to the study conducted by Koletsi et al(2012), it has been noted that there has been an increase in the percentage of articles accounting for the clustering effects in the statistical analysis. From only 25% of the included articles searched from 2010 and backwards to 51.9% of the included articles published in 2016 and 2017.

References:

1. Koletsi D, Pandis N, Polychronopoulou A, Eliades T. Does published orthodontic research account for clustering effects during statistical data analysis? European Journal of Orthodontics. 2012 Jun;34(3):287-92. PubMed PMID: 22015822. Epub 2011/10/22. eng.
2. Pandis N, Walsh T, Polychronopoulou A, Eliades T. Cluster randomized clinical trials in orthodontics: design, analysis and reporting issues. European Journal of Orthodontics 2013:669-675
3. Harrison JE, Burnside G. Why does clustering matter in orthodontic trials? European Journal of Orthodontics. 2012 Jun;34(3):293-5. PubMed PMID: 22628246. Epub 2012/05/26. eng.
4. Christie J, O'Halloran P, Stevenson M. Planning a cluster randomized controlled trial: methodological issues. Nursing Research. 2009 Mar-Apr;58(2):128-34. PubMed PMID: 19289934. Epub 2009/03/18. eng.
5. Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. BMC medical research methodology. 2004;4:21. PubMed PMID: 15310402. PMCID: PMC515302. Epub 2004/08/18. eng.
6. Kerry SM, Bland JM. Analysis of a trial randomised in clusters. British Medical Journal (Clinical research ed). 1998 Jan 3;316(7124):54. PubMed PMID: 9451271. PMCID: PMC2665333. Epub 1998/02/06. eng.
7. Burnside G, Pine CM, Williamson PR. Statistical aspects of design and analysis of clinical trials for the prevention of caries. Caries Research. 2006;40(5):360-5. PubMed PMID: 16946602. Epub 2006/09/02. eng.
8. Fleming PS, Koletsi D, Polychronopoulou A, Eliades T, Pandis N. Are clustering effects accounted for in statistical analysis in leading dental specialty journals? Journal of Dentistry. 2013 Mar;41(3):265-70. PubMed PMID: 23201411. Epub 2012/12/04. eng.
9. Altman DG, Bland JM. Statistics notes. Units of analysis. British Medical Journal (Clinical research ed). 1997 Jun 28;314(7098):1874. PubMed PMID: 9224131. PMCID: PMC2127005. Epub 1997/06/28. eng.
10. Bland JM, Kerry SM. Statistics notes. Trials randomised in clusters.British Medical Journal. (Clinical research ed). 1997;315(7108):600. PubMed PMID: 9302962.

11. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clinical trials (London, England). 2005;2(2):99-107. PubMed PMID: 16279131. Epub 2005/11/11. eng.
12. Campbell MK, Grimshaw JM. Cluster randomised trials: time for improvement. The implications of adopting a cluster design are still largely being ignored. British Medical Journal (Clinical research ed). 1998 Oct 31;317(7167):1171-2. PubMed PMID: 9794847. PMCID: PMC1114151. Epub 1998/10/31. eng.
13. Donner A. An empirical study of cluster randomization. International Journal of Epidemiology. 1982 Sep;11(3):283-6. PubMed PMID: 7129743. Epub 1982/09/01. eng.
14. Kerry SM, Bland JM. Sample size in cluster randomisation. British Medical Journal (Clinical research ed). 1998 Feb 14;316(7130):549. PubMed PMID: 9501723. PMCID: PMC2665662. Epub 1998/03/21. eng.
15. Gibson R, Harrison J. What are we reading? An analysis of the orthodontic literature 1999 to 2008. 2011. p. E471-E84.
16. Brignardello-Petersen R, Carrasco-Labra A, Booth HA, Glick M, Guyatt GH, Azarpazhooh A. A practical approach to evidence-based dentistry: How to search for evidence to inform clinical decisions. Journal of the American Dental Association (1939). 2014 Dec;145(12):1262-7. PubMed PMID: 25429040. Epub 2014/11/28. eng.
17. Hayes RJ, Moulton LH. Cluster randomised trials: Chapman and Hall/CRC; 2017.
18. Lawrence HP, Binguis D, Douglas J, McKeown L, Switzer B, Figueiredo R, Laporte A. A 2-year community-randomized controlled trial of fluoride varnish to prevent early childhood caries in Aboriginal children. Community Dentistry and Oral Epidemiology. 2008 Dec;36(6):503-16. PubMed PMID: 18422711. Epub 2008/04/22. eng.
19. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. Statistics in Medicine. 2012 Sep 10;31(20):2169-78. PubMed PMID: 22495809. Epub 2012/04/13. eng.
20. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. 2000.
21. Allan D. Some Aspects of the Design and Analysis of Cluster Randomization Trials. Journal of the Royal Statistical Society Series C (Applied Statistics). 1998;47(1):95. PubMed PMID: edsjsr.2986056.
22. Trutschel D, Palm R, Holle B, Simon M. Methodological approaches in analysing observational data: A practical example on how to address clustering and selection bias. International Journal of Nursing Studies. 2017 Nov;76:36-44. PubMed PMID: 28915416. Epub 2017/09/16. eng.

23. Papageorgiou SN. Handling qualitative research outcomes. Journal of Orthodontics. 2017 10 / 02 /;44(4):307-9. PubMed PMID: edselc.2-52.0-85032496704. English.
24. Bazargani F, Magnuson A, Lothgren H, Kowalczyk A. Orthodontic bonding with and without primer: a randomized controlled trial. European Journal of Orthodontics. 2016 Oct; 38(5) p. 503-7.
25. Qamruddin I, Alam MK, Fida M, Khan AG. Effect of a single dose of low-level laser therapy on spontaneous and chewing pain caused by elastomeric separators. American Journal of Orthodontics & Dentofacial Orthopedics. 2016 01/01/January 2016;149(1):62-6. PubMed PMID: S0889540615011063.
26. Mandall N, DiBiase A, Littlewood S, Nute S, Stivaros N, McDowall R, Shargill I, Wirthington H, Cousley R, Dyer F, Mattick R, Doherty B. Is early class III protraction facemask treatment effective? A multicentre, randomised, controlled trial: 15-month follow-up. 2010. Journal of Orthodontics 2010 Sep;37(3): 149-61
27. Tsiouli K, Topouzelis N, Papadopoulos MA, Gkantidis N. Perceived facial changes of Class II Division 1 patients with convex profiles after functional orthopedic treatment followed by fixed orthodontic appliances. American Journal of Orthodontics & Dentofacial Orthopedics. 2017 07/01/July 2017;152(1):80-91. PubMed PMID: S0889540617302640.
28. Mandall NA, Millett DT, Mattick CR, Hickman J, Macfarlane TV, Worthington HV. Adhesives for fixed orthodontic brackets. The Cochrane database of systematic reviews. 2003 (2):Cd002282. PubMed PMID: 12804432. Epub 2003/06/14. eng.
29. Wears RL. Advanced statistics: statistical methods for analyzing cluster and cluster-randomized data. Academic Emergency Medicine : official Journal of the Society for Academic Emergency Medicine. 2002 Apr;9(4):330-41. PubMed PMID: 11927463. Epub 2002/04/03. eng.
30. Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989. International Journal of Epidemiology. 1990 Dec;19(4):795-800. PubMed PMID: 2084005. Epub 1990/12/01. eng.
31. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Family practice. 2000 Apr;17(2):192-6. PubMed PMID: 10758085. Epub 2000/04/12. eng.
32. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. British Medical Journal (Clinical research ed). 1998 May 9;316(7142):1455. PubMed PMID: 9572764. PMCID: PMC1113123. Epub 1998/06/06. eng.

33. Kerry SM, Bland JM. Trials which randomize practices II: sample size. Family practice. 1998 Feb;15(1):84-7. PubMed PMID: 9527303. Epub 1998/04/04. eng.
34. Taljaard M, McRae AD, Weijer C, Bennett C, Dixon S, Taleban J, Skea Z, Eccles MP, Brehaut JC, Donner A, Saginur R, Boruch RF, Grimashaw J.M. Inadequate reporting of research ethics review and informed consent in cluster randomised trials: review of random sample of published trials. British Medical Journals (Clinical research ed). 2011 May 11;342:d2496. PubMed PMID: 21562003. PMCID: PMC3092521. Epub 2011/05/13. eng.
35. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. International Journal of Epidemiology. 2006 Oct;35(5):1292-300. PubMed PMID: 16943232. Epub 2006/09/01. eng.
36. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. International Journal of Epidemiology. 2015 Jun;44(3):1051-67. PubMed PMID: 26174515. PMCID: PMC4521133. Epub 2015/07/16. eng.
37. Manatunga AK, Hudgens MG, Chen S. Sample Size Estimation in Cluster Randomized Studies with Varying Cluster Size. Biometric Journal, 2001, 43(1): 75-86.
38. Kerry SM, Bland JM. Trials which randomize practices I: how should they be analysed? Family practice. 1998 Feb;15(1):80-3. PubMed PMID: 9527302. Epub 1998/04/04. eng.
39. Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. International Journal of epidemiology. 1999 april; 28(2):319-26. PubMed PMID: edsbas.CBD8BD19.
40. Edwards SJL, Braunholtz DA, Lilford RJ, Stevens AJ. Ethical issues in the design and conduct of cluster randomised controlled trials. British Medical Journal. 1999 02/10/accepted;318(7195):1407-9. PubMed PMID: PMC1115783.
42. McRae AD, Weijer C, Binik A, Grimshaw JM, Boruch R, Brehaut JC, Donner A, Eccles MP, Saginur R, white A, Taljard M. When is informed consent required in cluster randomized trials in health research? Trials. 2011 Sep 9;12:202. PubMed PMID: 21906277. PMCID: PMC3184061. Epub 2011/09/13. eng.
43. Jonathan ACS, George Davey S. Sifting The Evidence: What's Wrong With Significance Tests? British Medical Journal. 2001;322(7280):226. PubMed PMID: edsjsr.25226649.
44. Fisher RA. Book Reviews - Statistical Methods, Experimental Design, and Scientific Inference. A Re-issue of Statistical Methods for Research Workers, The Design of Experiments, and Statistical Methods and Scientific In. PubMed PMID: edsbas.22378E6A.
45. Fisher RA. The Design of experiments, 1922-1926, The American Statistician, 34 (1), 1-7

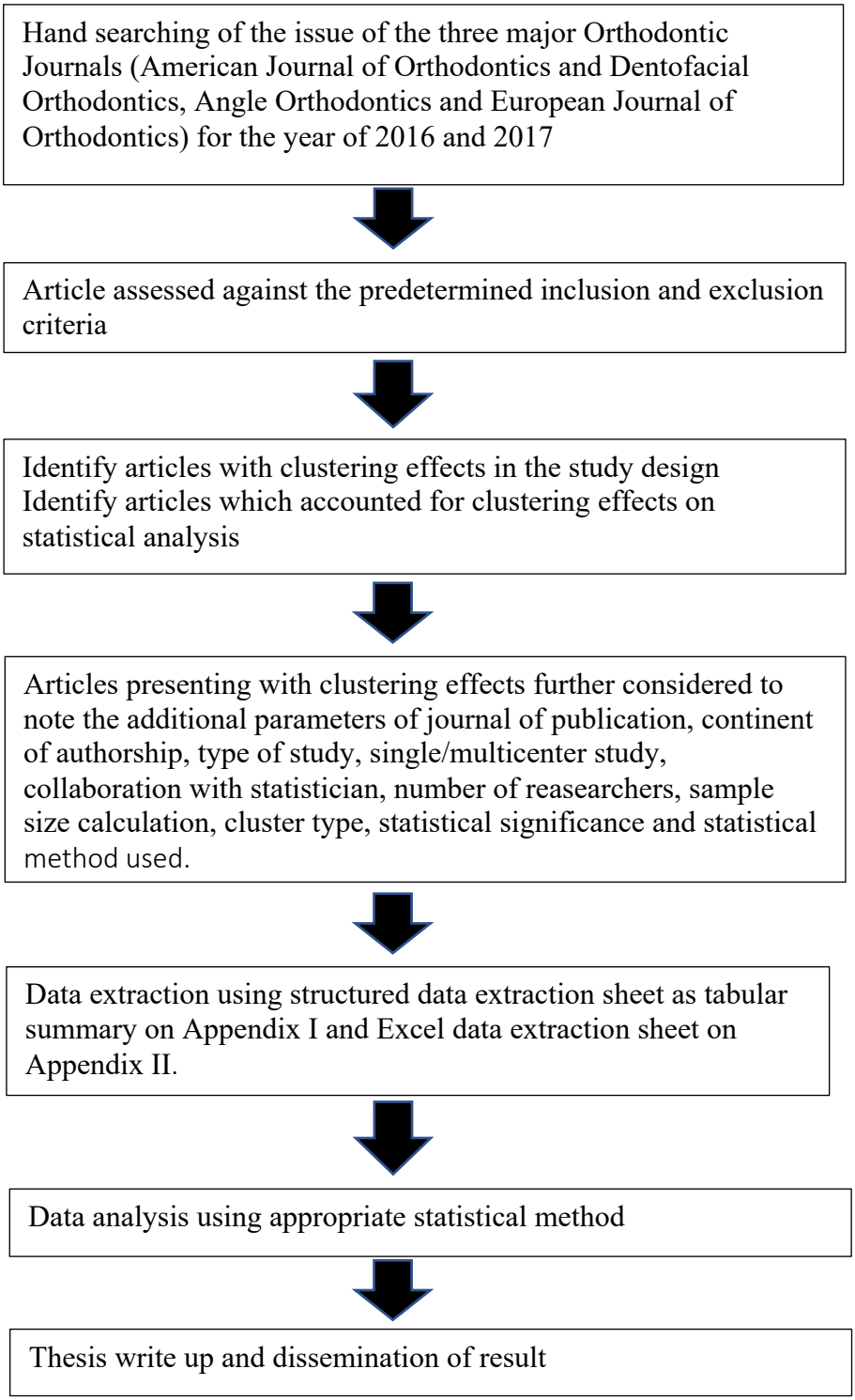
46. Ronald LW, Nicole AL. The ASA's Statement on p -Values: Context, Process, and Purpose. Germany, Europe: Figshare; 2016.
47. Dahiru T. P-Value, a true test of statistical significance? A cautionary note. *Annals of Ibadan Postgrauate Medicine*,2011, 6(1), 21-26. PubMed PMID: edsbas.A8836C1B.
48. Kim J, Bang H. Three common misuses of P values. Dental Hypotheses. 2016 07 / 01 /;7(3):73-80. PubMed PMID: edselc.2-52.0-84988734731. English.
49. Altman DG, Bland JM. Absence of evidence is not evidence of absence. Australian Veterinary Journal. 1996;74(4):311-. PubMed PMID: 8937675.
50. Kim H-Y. Statistical notes for clinical researchers: Type I and type II errors in statistical decision. RDE : Restorative Dentistry & Endodontics. 2015 (3):249. PubMed PMID: edskst.JAKO201513363799786:JAKO.
51. Chia KS. "Significant-itis"--an obsession with the P-value. Scandinavian Journal of Work, Environment & Health. 1997 Apr;23(2):152-4. PubMed PMID: 9167239. Epub 1997/04/01. eng.
52. Amitav B, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors.India Psychiatry Journal. 2009; 18(2): 127-131. PubMed PMID: edsbas.3D4EB660.
53. Pandis N. The P value problem. American Journal of Orthodontics and Dentofacial Orthopedics. 2013; 143(1), 150-151. PubMed PMID: edsbas.FFC1C460.
54. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal (Clinical Research ed). 1986;292(6522):746-50. PubMed PMID: 3082422. eng.
55. Nuzzo R. Scientific method: statistical errors. Nature. 2014;506(7487):150-2. PubMed PMID: 24522584.
56. Tan SH, Tan SB. The Correct Interpretation of Confidence Intervals. Proceedings of Singapore Healthcare. 2010 2010/09/01;19(3):276-8.
57. Polychronopoulou A, Pandis N, Eliades T. Appropriateness of reporting statistical results in orthodontics: the dominance of P values over confidence intervals. European Journal of Orthodontics. 2011 Feb;33(1):22-5. PubMed PMID: 20631084. Epub 2010/07/16. eng.
58. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology. 2016 April;31(4):337-50.
59. Savitz DA. Is statistical significance testing useful in interpreting data? Reproductive toxicology (Elmsford, NY). 1993;7(2):95-100. PubMed PMID: 8499671. Epub 1993/01/01. eng.

60. Kirkwood B R SJAC. Essential Medical Statistics. 2nd edition ed: Blackwell Publishing, Oxford; 2003.
61. Walenkamp MMJ, Roes KCB, Bhandari M, Goslings JC, Schep NWL. Multiple testing in orthopedic literature: a common problem? BMC research notes. 2013;6:374. PubMed PMID: 24053281.
62. Sainani KL. The problem of multiple testing. PM R. The Journal of Injury, Function, and Rehabilitation. 2009;1(12):1098-103. PubMed PMID: 20006317.
63. Gordi T, Khamis H. Simple solution to a common statistical problem: Interpreting multiple tests. Clinical Therapeutics. 2004 2004/05/01/;26(5):780-6.
64. Koletsi D, Karagianni A, Pandis N, Makou M, Polychronopoulou A, Eliades T. Are studies reporting significant results more likely to be published? American Journal Orthodontics Dentofacial Orthopedics. 2009; 136(5): 632. e1-633.
65. Papageorgiou SN. Tooth-level versus patient-level. Journal of Orthodontics. 2018; 01 / 02 /;45(1):51-53. PubMed PMID: edselc.2-52.0-85042224001. English.
66. Manning N, Chadwick SM, Plunkett D, Macfarlane TV. A randomized clinical trial comparing `one-step' and `two-step' orthodontic bonding systems. Journal of Orthodontics. 2006; 33(4): 276-257.
67. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. British Medical Journal (Clinical research ed). 2004 Mar 20;328(7441):702-8. PubMed PMID: 15031246. PMCID: PMC381234. Epub 2004/03/20. eng.
68. Jia B, Lynn HS. A sample size planning approach that considers both statistical significance and clinical significance. 2015. PubMed PMID: edsbas.FDB903B5.
69. Shimada T, Takayama H, Nakamura Y. Quantity and quality assessment of randomized controlled trials on orthodontic practice in pubmed. Angle Orthodontics. 2010;80(4) 525-530.
70. Papageorgiou SN, Eliades T, Antonoglou GN, Martin C. Methods, transparency and reporting of clinical trials in orthodontics and periodontics. Journal of Orthodontics. 2019 06//;46(2):101.
71. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. Trials. 2010 03 / 24 /;11. PubMed PMID: edselc.2-52.0-77952279621. English.
72. Koletsi D, Pandis N, Fleming PS. Sample size in orthodontic randomized controlled trials: are numbers justified? European Journal of Orthodontics 2014; 36(1):67-73.

73. Pandis N, Polychronopoulou A, Madianos P, Makou M, Eliades T. Reporting of Research Quality Characteristics of Studies Published in 6 Major Clinical Dental Specialty Journals. Journal of Evidence Based Dental Practice. 2011/06/01/;11(2):75-83.
74. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clinical trials (London, England). 2004;1(1):80-90. PubMed PMID: 16281464.
75. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet (London, England). 1997;350(9074):326-9. PubMed PMID: 9251637.
76. Harrison J. Evidence-based orthodontics--how do I assess the evidence? Journal of Orthodontics, 2000; 27(2), 189-197 PubMed PMID: edsbas.78DB787E.

Appendices

Appendix I: Methodology Flow Chart



Appendix II: Tabular summary on the total number of articles in Word format

Journal & Volume of publication	Issue and month of publication	Number of articles assessed	Number of articles excluded	Number of articles considered for clustering	Number of articles considered to have clustering effects	Number of articles accounted for clustering in statistical analysis	Number of articles did not account for clustering in statistical analysis	Number of articles with outcomes assessed separately
TOTAL								

Appendix III: Excel Data Extraction form

No	Journal of Publication	Title of article	Are clustering effects accounted for	Study type	Continent of authorhship	Number of researchers	Involvement of statistician	Statistical significance	Statistical method used	Is the sample size reported	Remarks